Lappeenranta-Lahti University of Technology LUT
School of Business and Management
Master's Programme in Strategic Finance and Business Analytics

Master's thesis

Supervised feature selection methods for default prediction in P2P lending

Juhana Hautakangas

2020

1st examiner: Christoph Lohrmann

2nd examiner: Mikael Collan

ABSTRACT

Author:	Juhana Hautakangas
Title:	Supervised feature selection methods for default prediction in P2P lending
Faculty:	LUT School of Business and Management
Master's programme:	Master's Programme in Strategic Finance and Business Analytics
Year:	2020
Master's Thesis:	99 pages, 14 appendices, 19 tables, 24 figures
Examiners:	Postdoctoral Researcher Christoph Lohrmann and Professor Mikael Collan
Keywords:	P2P lending, feature selection, default prediction

The purpose of this thesis is to investigate the performance of different feature selection (FS) methods in P2P lending default prediction. The tested FS methods include maximum-relevance-minimum-redundancy (MRMR) approach, Chi-Square FS method, sequential forward selection (SFS) method and learning model-based feature ranking (LMBRF) method. The FS methods are examined in combination with Naïve Bayes (NB), logistic regression (LR), decision tree (DT) and random forest (RF) classifiers. A systematic comparison of the used models is conducted using historical P2P loan data provided by Bondora, an Estonian P2P lending platform.

The performance of FS methods is evaluated based on the final classification performance and model complexity. Classification performance is measured using both the performance metrics calculated based on the confusion matrices and the area under the ROC curve (AUC) metric. The model complexity is measured by the number of used features in the final classification models.

The study results indicate that all the tested FS methods are suitable for FS in P2P lending default prediction context. Using each of the FS methods, at least competitive classification performance was obtained compared to the models without FS, with considerably smaller number of features. Overall, the SFS method was found to be the most efficient of tested FS models. It was the only method that managed to improve the classification accuracy statistically significantly with almost all the tested classification models and it also helped to reduce the number of features most considerably. Other investigated FS methods were found to perform somewhat equally compared to each other.

TIIVISTELMÄ

Tekijä:	Juhana Hautakangas
Otsikko:	Ohjatut muuttujanvalintamallit vertaislainojen luottoriskin ennustuksessa
Akateeminen yksikkö:	LUT School of Business and Management
Maisteriohjelma:	Master's Programme in Strategic Finance and Business Analytics
Vuosi:	2020
Pro gradu:	99 sivua, 14 liitettä, 19 taulukkoa, 24 kuviota
Ohjaajat:	Tutkijatohtori Christoph Lohrmann ja professori Mikael Collan
Hakusanat:	Vertaislainaus, muuttujanvalinta, luottoriskin ennustus

Tämän tutkielman tavoitteena on tutkia erilaisten muuttujanvalintamallien suoriutumista vertaislainojen luottoriskin ennustuksessa. Tutkittavina muuttujanvalintamenetelminä käytetään MRMR (maximum-relevance-minimum-redundancy) -menetelmää, khiin neliö -testiin perustuvaa menetelmää, eteenpäin askeltavaa muuttujanvalintamallia sekä luokittelumalleihin pohjautuvaa muuttujien järjestämistä. Valintamalleja arvioidaan käyttämällä niitä yhdessä koneoppimiseen perustuvien luokittelumallien kanssa. Luokittelumalleina käytetään naiivia Bayes -luokittelijaa, logistista regressiota, päätöspuita ja satunnaisia metsiä. Tutkimusaineistona hyödynnetään virolaisen vertaislaina-alustan Bondoran historiallista lainadataa.

Muuttujanvalintamallien suoriutumista arvioidaan ennustusmallien lopullisen luokittelutehokkuuden sekä mallien monimutkaisuuden perusteella. Luokittelutehokkuutta mitataan käyttämällä erilaisia sekaannusmatriisiin perustuvia tunnuslukuja sekä AUC (area under the ROC curve) -tunnuslukua. Mallien monimutkaisuutta arvioidaan lopullisissa luokittelumalleissa käytettyjen muuttujien lukumäärän perusteella.

Tutkimustulokset osoittavat, että kaikki testatut muuttujanvalintamallit soveltuvat käytettäväksi vertaislainojen luottoriskin ennustuksessa. Tulosten mukaan kaikkien muuttujanvalintamallien hyödyntäminen johti vähintään kilpailukykyiseen luokittelutehokkuuteen verrattuna malleihin ilman muuttujanvalintaa, selkeästi pienemmällä muuttujamäärällä. Tutkituista malleista tehokkain oli eteenpäin askeltava muuttujanvalintamalli, joka tutkituista malleista ainoana paransi luokittelutehokkuutta tilastollisesti merkitsevästi lähes kaikkien luokittelumallien kohdalla. Kyseisen muuttujanvalintamallin avulla myös muuttujamäärää onnistuttiin vähentämään merkittävimmin. Muut muuttujanvalintamallit olivat tehokkuudeltaan keskenään jokseenkin tasavertaisia.

ACKNOWLEDGEMENTS

The last five years as a full-time student in LUT have been a very special period in my life. These years in Lappeenranta have offered memorable experiences, long days of hard work and nice time with new and old friends. Now, at the end of this journey, I cannot even realize it is over. It is time to take a step towards unknown and move towards new challenges.

I want to give a very special thanks to my supervisor Christoph Lohrmann for professional advice through the whole thesis process. Your professionality was essential for compliting the thesis, and I really appreciate the effort you put into answering all my emails and making careful suggestions for improvement. A big thanks also to Mikael Collan for suggesting the interesting research topic and guiding me through the thesis process.

Thanks to my family for all the caring and encouragement I have received from you through my studies, your support has been irreplacable. Special thanks to my fionce Saara for always pushing me forward towards my dreams. Without your endless support I wouldn't have made it through this journey. I cannot thank you enough for being loving and patient also in my worst days during the years. Thanks also for encouraging me to apply for the school for another time during my military service.

Thanks also to all the people in Fazer Lappeenranta for your encouragement through the years and for understanding the challenges in combining the work and studies. Thanks for providing a possibility to finance my studies through summer and weekend jobs.

Special thanks to all the old and new friends for your support during these years. Especially, all the people I have met through floorball over these years deserve praise for good moments and memories. Last but not least, huge thanks to Tuomas for all the support I have received from you through these years. Without your endless encouragement and unselfish help this school journey would have been much more painful. I really appreciate your kindness and all the moments we have spent together – in school and on free time.

In Lappeenranta, June 18th, 2020

Juhana Hautakangas

TABLE OF CONTENTS

TABLES

FIGURES

APPENDICES

LIST OF ABBREVIATIONS

AUC	Area under the ROC curve
CV	Cross-validation
DT	Decision tree
(L)DA	(Linear) discriminant analysis
FN	False negatives
FP	False positives
FS	Feature selection
GA	Genetic algorithm
GP	Genetic programming
KNN	K-nearest neighbor
LMBFR	Learning model-based feature ranking
LR	Logistic regression
MARS	Multivariate adaptive regression spline
MDA	Mean decrease accuracy
MDI	Mean decrease impurity
MI	Mutual information
ML	Machine learning
MRMR	Maximum-relevance-minimum-redundancy
(A)NN	Artificial neural network
P2P	Peer-to-peer
RF	Random forest
ROC	Receiver operating characteristic curve
SVM	Support vector machines
S(F)FS	Sequential (floating) forward selection
TN	True negatives
TP	True positives

1 INTRODUCTION

The rapid evolution and growing popularity of internet and online communities have considerably changed the world during the last decades, and the financial sector has not been left out of this development. *Peer-to-peer (P2P) lending* is a good example related to the structural changes happening in the financial industry (Berger and Gleisner 2009). It is a relatively new lending model in which borrowers and lenders are matched directly through an online platform without a financial institution acting as an intermediary. P2P lending has rapidly gained popularity in recent years especially because of its flexibility and relatively high returns on investment (Bachmann et al. 2011).

While P2P lending platforms have become more popular, also the related problems such as fraud and incompetence have caused more and more debate (Chen et al. 2014). Credit risk management is facing new challenges in the context of social online lending because the P2P loans are unsecured, and the platforms are relatively heterogenous. One of the most common research topics related to P2P lending has been credit scoring and the identification of borrowers that are more likely to default than others. Besides that, the detection of successful borrowers has frequently been under consideration in previous studies. Many methods, including different machine learning algorithms and data mining techniques have been developed to predict the P2P lending default (Eunkyoung and Lee 2012). However, no consensus still exists regarding the most accurate default prediction model in P2P lending context.

As many real-world datasets, P2P loan datasets are typically large and high-dimensional: they usually cover a lot of observations and *features* (also referred to as *predictors*, *input variables* or *independent variables*). Different automized predictive algorithms are often applied to make use of this kind of data. Because the datasets can have a large number of dimensions, the models frequently become complex and computationally expensive. Irrelevant variables also introduce excess noise into the models which decreases their performance (Dash and Liu 1997). To solve these challenges, different dimensionality reduction methods are used.

Feature selection (FS) is a dimensionality reduction technique which is used to select the most appropriate subset of the set of all available features (Dash and Liu 1997). Through the FS, the irrelevant and redundant features can be removed from the data supplied to the predictive models and more attention can be paid to the most relevant variables. FS can frequently reduce overfitting, improve predictive performance, and decrease the computational costs of the used prediction models (Guyon and Elisseeff 2003).

1.1 Motivation and background

P2P lending platforms typically attract investors by advertising their relatively high returns compared to the traditional investment products. For example, a US P2P lending platform Lending Club reported approximately 13% average annual return for the investors on the last quarter of 2019, while the interest rates have been historically low in recent years (Lending Club 2019). However, also the risks associated with P2P lending are relatively high for example due to the problems of information asymmetry and the unsecured nature of P2P loans (Chen et al. 2014). Therefore, the lenders must have tools to analyze the creditworthiness of the borrowers and to discriminate the investments that are attractive from a risk-reward standpoint from the ones that are not worth investing in.

Many P2P lending platforms provide historical data of loans that have been made through them. This gives an opportunity for the investors to make their own investment analysis based on this data. It has also made it possible to research the special features of the new form of financing. The P2P loan datasets usually contain a large number of variables, typically providing information for instance about the borrower's demographics and the characteristics of the loan. Different *machine learning* (ML) algorithms are often used to analyze such complex data, and ML models are widely applied also in P2P lending default prediction (Berger and Gleisner 2009).

A ML model uses a large quantity of past input data to learn the underlying structure and patterns of the data. The model is trained with the past data, and then the trained model is used to predict the values of the target variable of the new, unseen data (Bishop 2006, pp.3-4). The complexity of ML models increases markedly when the number of dimensions grows, and different kind of FS methods are used to automatize the removal of irrelevant and redundant variables from the models (Dash and Liu 1997).

In the previous research focusing on the default prediction, the FS has often relied on the intuition, earlier knowledge of the field and usually unsystematic arbitrary trial (Liu and Schumann 2005). However, in the complicated datasets, the inter-relationships between variables can be unexpected. This is especially the case when dealing with the datasets that are affected by human behavior, often characterized by limited rationality. Therefore, the selection of the most relevant features based on the intuition can lead to the removal of features that are, in reality, significant for the default prediction. In this study, the performance of different automized FS methods is tested in P2P lending default prediction area.

A systematic performance comparison of different FS methods in P2P lending default prediction has not been done earlier (at least to best of my knowledge), which serves as a research

gap for this study. The results of this study can be exploited by investors when making the investment decision on the P2P platforms. The findings can help the investors to construct more accurate models to predict the default of the loans and to discriminate between bad and good loan applicants.

1.2 Focus of the study

The focus of this study is on the use of FS methods in P2P lending default prediction. Figure 1 represents the most important concepts related to the area of the research subject and their inter-relations. FS is an essential part of ML: it plays an important role in the construction phase of ML-based forecasting and classification models. ML, in turn, is related to the credit risk management because the ML models are often used as tools in credit risk assessment and credit scoring. Finally, the data and motivation for this study come from the P2P lending area.

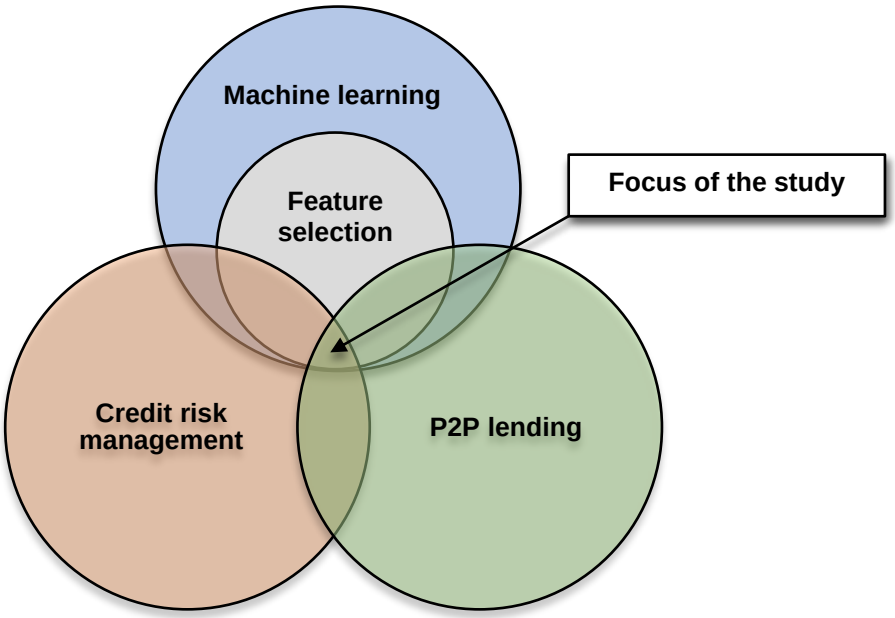


Figure 1. Focus of the study

The main focus of the study is represented in the intersection of four circles: it combines all the four above-mentioned concepts. The theoretical framework of the study is based on the literature related to these fields of research.

1.3 Research questions and limitations

The main objective of this thesis is to compare the performance of different FS methods in P2P lending default prediction. To make the comparison as reliable as possible and to build accurate classification models for predicting the default, it is important to examine the previous

scientific literature and empirical studies about the subject. The literature review of this study examines the previous research related to the topic, and the first research question and three related sub-questions are formed as follows:

1. What is the current state of credit risk assessment and prediction in the scientific literature?
    a. What statistical and machine learning models have been the most popular in credit risk assessment and prediction in previous studies?
    b. How has the feature selection been performed in previous studies and how have the methods been evaluated?
    c. What variables have been found to explain the credit risk in P2P lending in previous studies?

The second research question is related to the performance of different FS methods and is the main research question of this study. It is formed as follows:

2. How do different feature selection methods perform compared to each other in P2P lending default prediction?

To answer the research question, different FS methods are used to select the features for different classification models which are used to predict the loan default. The performance of different FS methods is tested using historical loan data provided by an Estonian P2P platform *Bondora* (introduced further in Chapter 2). The performance comparison is made by comparing the final classification performance of different classifiers to each other using the feature subsets proposed by different FS methods. Also, as proposed by Liu and Yu (2005), a simple "before-and-after" experiment is conducted. In this approach, the classification accuracy obtained using the full set of features is compared to the accuracy obtained using the feature subsets proposed by different FS methods. In addition to the classification performance, model complexity (the number of used features in the final models) is also considered in the comparison.

In addition to comparing the performance of different FS methods, this study aims to investigate the important features in discriminating the default loans from non-default loans in Bondora data. Therefore, the third research question is formed as follows:

3. What are the most important features in predicting the default in Bondora dataset?

The research questions are answered during this thesis. No consensus exists among researchers regarding either the most efficient FS methods or the most important features

determining the P2P lending default risk. The results are also dependent on the used data and models. For these reasons, no research hypotheses are formed at this point of the expected results.

The study is limited to focus on the P2P lending default prediction because the use of different FS methods in this area is scarce and needs more attention. Furthermore, the study results are based on a single dataset provided by one P2P lending platform. This delimitation of data limits the potential generalization of the results but is vital to effectively investigate the scope of the thesis within given time and length limits.

1.4 Structure of the thesis

The thesis consists of 7 chapters and begins with a brief introduction. The structure of the thesis after the introduction is illustrated in Figure 2. Overall, the thesis can be divided into two main parts: the first part focuses on the theoretical aspects of the topic, and in the second part, the empirical analysis is conducted. After the introduction, the theoretical framework of the topic is introduced (the principles of P2P lending, ML-based prediction and FS are described), and the literature review of previous research is conducted.

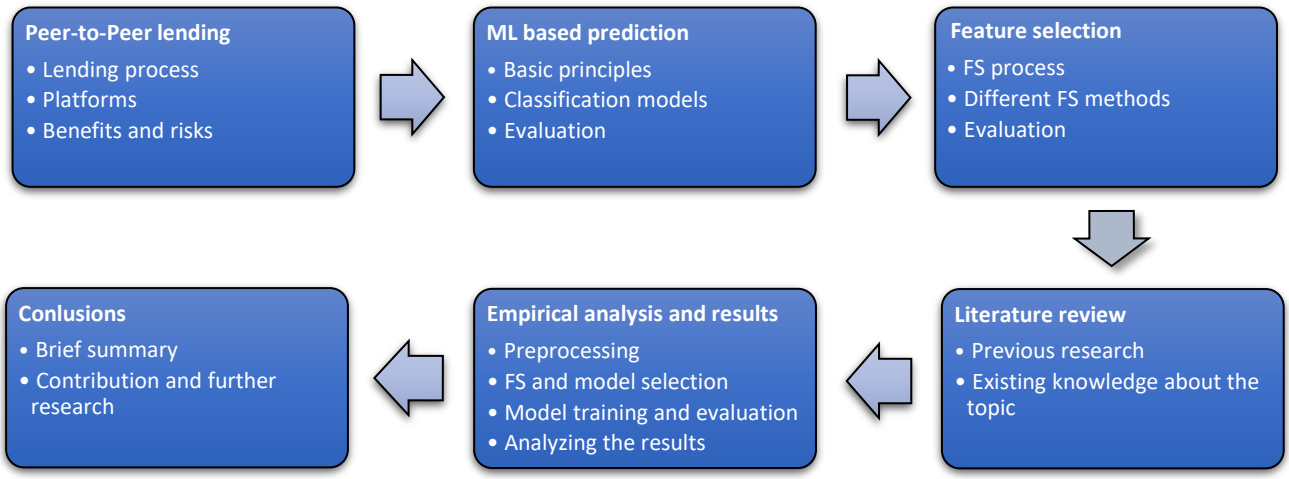


Figure 2. Structure of the thesis

In the empirical part of the thesis, the actual empirical analysis is reported. First, the used data and the pre-processing steps are described. Then, different combinations of ML algorithms and FS methods are used to predict the default on P2P lending dataset. After that, the results are introduced and discussed. Finally, the conclusions are made based on the obtained results and the contribution of the study for the financial research is discussed.

2 PEER-TO-PEER LENDING

One of the most significant causes of the latest financial crisis was the long-standing deregulation in the financial and banking sector. To solve the structural problems behind the crisis and to restore the confidence in the banking sector, the regulation of the financial industry was tightened considerably after the crisis. This has led to a situation where borrowing from the banks and other traditional lenders has become difficult or even impossible for the borrowers with low credit ratings because traditional lenders have refrained from high-risk lending. At the same time, low interest rates maintained by central banks have reduced the returns on savings and investments of the households and investors (Crotty 2009).

P2P lending can be considered as a potential answer for these problems. It can be defined as a practice of lending money to people or companies through online platforms that match lenders and borrowers directly without a financial intermediary (Zhao et al. 2017). P2P lending platforms offer an alternative way to borrow funds for private individuals and businesses which cannot borrow from the traditional lenders or are seeking for better loan conditions. In return, they give an opportunity for investors to achieve relatively high returns on their investments. In P2P lending markets, the investors are typically private individuals (non-professional investors) which frequently strongly affects the investment behavior (Bachmann et al. 2011).

As stated by Berger and Gleisner (2009), another reason for the rapid growth of P2P lending has been the rapid development of information technology and online communities in recent years. This has led to the evolution of new electronic marketplaces where the role of traditional intermediaries (banks and other financial institutions) has been decreased considerably or even completely eliminated. New forms of electronic lending compete with traditional bank lending for example with smaller fixed costs and the easiness of the lending process. However, there are also significant drawbacks related to P2P lending which are related to higher risks associated with the new lending model (Yum et al. 2012).

In this chapter, the P2P lending process is first explained briefly to get an overall look of the procedure. After that, a few established P2P platforms are introduced briefly, and the biggest benefits and risks of P2P lending are discussed. Finally, the credit risk assessment and management in P2P lending context are considered.

2.1 P2P lending process

Even if the developed P2P lending platforms differ from each other in many ways, the main process is usually relatively similar across different platforms. The borrowing process begins with the registration phase in which borrowers register to the platform and give personal

information about themselves. During the registration process, the identity of the borrower is strictly verified by requiring private information such as ID card number, and the registration form typically secures professional, personal and financial details of the borrower. Usually, some kind of credit rating is assigned to the borrower based on the information given on the registration form (Wang et al. 2015). Some platforms limit the registration to certain groups of people, for example in a large US P2P lending platform Lending Club both the investors and borrowers are required to be US residents (Zhao et al. 2017).

After the registration, the borrowers fill out the actual loan application in which they determine the amount of money they want to borrow and the maximum interest rate they are willing to pay. Some other information about the loan is also typically required (or given optionally), such as the use of the loan, repayment period and monthly cost. When the loan application has been filled out, the loan is listed for potential lenders (Wang et al. 2015).

As well as the borrowing process, a typical P2P lending (or investment) process also begins with the registration into the chosen P2P lending platform. When the registration phase is completed, the lender decides the amount and the time period of an investment. After that, the search process of potential loans (and borrowers) takes place. The search of potential loans (or borrowers) is done either manually by the lender or automatically by the platform. Usually, the investor does the investment decision based on the information provided by the borrower (Klafft 2008; Wang et al. 2015).

As stated by Wang et al. (2015), there are two popular ways to make an investment on the P2P lending platform: in the first model, the lender chooses the borrower from the platform by himself and lends the money directly to the borrower. In the second model, the lender invests in a pool of funds which matches his desired risk category and loan maturity and the money is allocated to the corresponding borrowers by the platform. The drawback of the second option is that the lender does not have individual information of the borrowers. Contrarily, in case of the first option, the manual search of potential borrowers can be time consuming (Davis and Murphy 2016; Wang et al. 2015).

When the loan request of the borrower is fully funded by the lenders through the lending process, many platforms require another verification of borrower's repayment ability, usually including the verification of steady income of the borrower (Bachmann et al. 2011). Finally, if all the verifications are fulfilled, money is transferred from the lender's account to the borrower's account. After that, the borrower begins the repayment process according to the negotiated schedule. The loan request can be fully funded by an individual investor but frequently the loan applications have multiple investors (Bachmann et al. 2011; Wang et al. 2015).

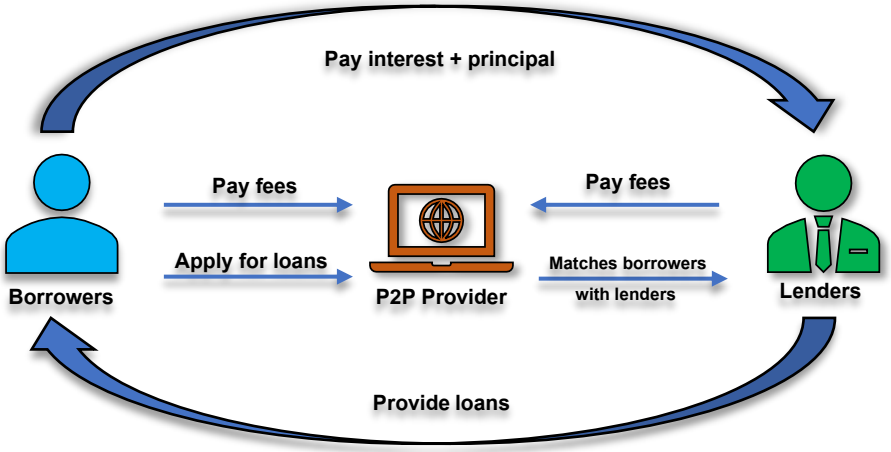


Figure 3. Simplified illustration of P2P lending process

The simplified illustration of a P2P lending process is represented in Figure 3. First, the borrower applies for a loan through the P2P lending platform with certain conditions. Then, the P2P provider matches the borrower (loan) with a potential lender, and the lender provides a loan for the chosen borrower. The loan is granted under the loan conditions accepted by opposite parties (borrower and lender), and the borrower pays the interest and the principal of the loan back to the lender under the loan period. In addition, both the borrower and the lender pay fees to the P2P provider for providing the service (Davis and Murphy 2016). For example, a US P2P platform Lending Club charges a loan origination fee from borrower which ranges from 1% to 6% of the loan amount, based on the borrower's credit rating. The same platform charges a service fee from lender which equals approximately 1% of the amount of each payment made by the borrower (Berger and Gleisner 2009; Lending Club 2020a).

2.2 P2P lending platforms

P2P lending platforms vary from each other in many ways. Among the biggest differences between platforms are the differences in pricing mechanisms. In a commonly used *posted price* mechanism the platform determines the interest rates of the loans according to the expected creditworthiness of the borrower and the loan conditions. Instead, in an *auction-based* mechanism the price (interest rate) of the loan is determined by the lenders with an auction process. Additionally, in many platforms the borrower sets the maximum interest rate he is willing to pay, and the lenders decide whether they want to invest on the loan with the given rate (Wei and Lin 2017). Table 1 lists examples of several P2P platforms and basic information about them. The listed platforms are well-known and established P2P lending providers and represent different major P2P lending market regions: US, China and Europe.

Table 1. Examples of P2P lending platforms

Platform	Zopa	Prosper	Lending Club	PPDai	Bondora
Home country	United Kingdom	United States	United States	China	Estonia
Founded	2005	2005	2006	2007	2009
Pricing mechanism	Initially auction, now posted prices	Initially auction, now posted prices	Posted prices	Auction	Borrower's maximum interest rate
Currency	British pound	US dollar	US dollar	Chinese yuan	Euro
Loan amount	£1000 - £25000	$2000 – $35 000	$1000 – $40 000	¥100 – ¥200 000	500€ – 10 000€
Loan term	1 to 5 years	3 or 5 years	3 or 5 years	1 to 24 months	3 to 60 months
Cumulative loan amount (through lifetime)	About £5.3 billion (March 2020)	About $16.7 billion (March 2020)	About $56.8 billion (March 2020)	Not available*	About 370 million € (March 2020)

*The cumulative loan amount through lifetime not available. The loan origination volume was about ¥24.6 billion during the 3rd quarter of 2019.

Zopa, founded in United Kingdom in 2005, was the first P2P platform in the world. It started as an auction-based platform but nowadays it offers four loan products in which the loans are diversified based on their riskiness. In Zopa, the invested money is automatically divided across multiple borrowers and therefore the risk of the investment is always diversified. Zopa has a good reputation among P2P platforms due to its relatively low default rates. The amount of loans made through the platform is about 5.3 billion British pounds (Zhao et al. 2017; P2PMarketData 2020).

Prosper was founded in 2005 and was the first American P2P lending platform. It started as an auction-based platform as well but switched to posted price mechanism in 2010. The cumulative amount of loans made through the platform is about 16.7 billion US dollars (P2PMarketData 2020; Wei and Lin 2017; Zhao et al. 2017). Prosper was one of the first P2P platforms that made their loan data publicly available for their users and, therefore, the data has been used in many empirical studies (Iyer et al. 2009; Guo et al. 2016).

Lending Club is nowadays the world's largest online P2P lending platform. The cumulative amount of loans originated through the platform is about 56.8 billion US dollars (P2PMarketData 2020). It uses the posted price mechanism: the interest rate is assigned to each loan according to the loan grade which is determined by the risk level of the loan and the creditworthiness of the borrower. Lending Club also provides the historical data of loan applications for its users, and the Lending Club data has been widely used in empirical studies in P2P lending area (Zhao et al. 2017).

PPDai was the first P2P platform in China. It uses an auction-based pricing mechanism and the loan origination volume through the platform during the 3rd quarter of 2019 was about 24.6 billion Chinese yuans. PPDai changed its legal name to "FinVolution Group" in 2019 but is still

more commonly known as “PPDai” or “Paipaidai” (Finvolution Group 2019; Yuang and Wang 2016, pp.66-67). PPDai loan data has been used in many empirical studies concerning the Chinese P2P lending markets (Chen 2019; Zhang 2016).

The data for this study is provided by *Bondora* which is an Estonian P2P lending platform that started operating in 2009. Ever since, almost 120 000 people have invested altogether almost 370 million euros through the platform. Bondora has its focus on the unsecured consumer loans in which the principal amounts are between 500€ and 10 000€. The pay-back periods of the loans range from 3 to 60 months. The loan appliers are mostly Estonian, Finnish, Spanish or Slovakian, but the platform has investors from 40 countries (Bondora 2017; P2PMarketData 2020).

2.3 Benefits of P2P lending

P2P lending has many benefits compared to the traditional lending models. As described earlier, the platforms offer the possibility to get funded for borrowers who do not have access to bank loans. Borrowers can also frequently borrow money with better loan terms than in the case of traditional lending. This is due the low cost structure of P2P lending which can be explained by relatively small overhead costs: the whole P2P lending process is done through the online platform and therefore the operational costs are lower compared to the costs of traditional banks (Pokorna and Sponer 2016; Zhao et al. 2017).

Other benefits of P2P lending include for instance the increased flexibility and easiness of the loan application process. Because the loan application is filled out online, it can be done independently of place and time. In addition, the approval process is usually fast and easy: the funding decision is typically made much faster than in the case of traditional borrowing process. Also, in contrast to traditional lending, the loan conditions usually do not include any requirements of collateral. Furthermore, the loan conditions can also be better tailored according to the preferences of the borrower (Pokorna and Sponer 2016).

On the investors’ side, the platforms typically offer more attractive returns than the traditional investments. According to Pokorna and Sponer (2016), the P2P lending platforms have provided above 10% annual return for investors during the years of very low interest rates after the latest financial crisis. P2P lending also offers the possibilities of diversification for the investors who invest mostly in traditional investments. Also, the diversification on the P2P platform itself is easy: the invested amount can be easily divided between multiple loans (Wang et al. 2015). In P2P lending, the elimination of expensive intermediaries also reduces the transaction costs. The lending process is also transparent because the lenders typically choose the

borrowers by themselves and wide background information about the borrowers is often available (Klafft 2008).

2.4 Risks of P2P lending

In contrast to the obvious benefits of P2P lending compared to the traditional lending models, the risks associated with the P2P loans are also considered high. The P2P providers do not typically carry the credit risk, but it is left to the lenders. The event of default in P2P lending context typically leads to at least a partial loss of the loan amount and interest payments because the P2P platforms typically do not guarantee the loan payback and the loan conditions usually do not require collaterals (Pokorna and Sponer 2016). The investors are often non-professional and therefore frequently do not have enough financial expertise to comprehensively assess the risks of the investments even though the required information would be available (Klafft 2008). In addition, despite the fact that the P2P platforms have developed different ways to confirm the borrower information, it is possible that the borrowers misrepresent the information about their creditworthiness (Pokorna and Sponer 2016).

Yum et al. (2012) claim that the information asymmetry is one of the most significant fundamental problems faced by the P2P platforms. The pseudonymous nature of P2P lending platforms increases the risk of borrowers' opportunistic behavior at the expense of lenders and worsens the problems of adverse selection. This leads to the situation where people with higher risk of default are more willing to borrow money from the P2P lending platforms than the people who are expected to pay their loans back successfully. It is noteworthy that a big part of P2P borrowers does not have access to the traditional bank loans due to the low credit rating (Yum et al. 2012).

In addition, P2P lenders are also exposed to the agency risk. In the P2P lending context, the agency risk is related to the possibility that the platform goes bankrupt or ceases its operations because of the unprofitability of the business. Also, the failures of the platform software can lead to losses for the investors (Davis and Murphy 2016).

Furthermore, the P2P lending markets are frequently characterized by illiquidity of investments (Davis and Murphy 2016). The maturities of the loans can be relatively long and for example Bondora offers the loans with the loan period up to 60 months (Bondora 2017). To reduce the illiquidity problems, many P2P platforms have developed secondary markets for their loans. However, there are still P2P platforms in which selling the loans on the secondary market is not possible or whose secondary markets suffer from bad efficiency. In the inefficient secondary market, there might not be enough buyers and sellers for the loans which leads to illiquidity

problems: the investors willing to sell their loans might not be able to do that due to the absence of interested buyers (Pokorna and Sponer 2016).

Another issue that is commonly considered regarding the P2P lending is the relatively scarce regulation of the field. While the regulation of financial industry in general has been tightened notably in recent years, the P2P lending markets are still underregulated to some extent. The regulation typically does not insure the investments in P2P lending platforms (P2P loans are not covered by the deposit insurance) even though many P2P platforms offer (by charge) their own buyback guarantees. Also, because of unique characteristics of P2P lending market, the P2P lending operators do not have to restrict themselves according to the bank regulations such as Basel III capital and liquidity requirements. However, it is worth mentioning that most of the P2P platforms hold reserve funds to compensate the losses of investors if needed (Davis and Murphy 2016; Pokorna and Sponer 2016).

As Davis and Murphy (2016) state, the regulators around the world have recently noticed the need for the legislation of growing P2P lending markets. A topical example of building the regulation concerning the P2P lending can be mentioned from Finland. In Finland, the new consumer protection law came into effect on September 1st of 2019 which also affects the P2P lending market. The purpose of the new law is to reduce the growing indebtedness and increase the transparency in consumer lending industry. It caps the interest rates of all unsecured loans (including P2P loans) to 20% p.a. and sets some limits for the costs related to consumer loans (Yle 2019).

2.5 Assessing and managing credit risk in P2P lending

Due to the high-risk nature of the P2P lending, from investor's point of view it is essential to assess and manage the credit risk efficiently. Different ways have been developed to conduct the credit risk assessment and management in P2P lending context (Pokorna and Sponer 2016).

2.5.1 Credit scoring systems of P2P platforms

The P2P platforms typically provide their own credit scoring estimates for the loan applications. These credit ratings are based on the financial and personal information which the borrower has given when completing the registration and the loan application process. The credit ratings are typically derived based on the statistical techniques exploiting the historical records of loan applicants and their repayment ability. These "in-house" credit ratings are frequently supported by the credit ratings obtained from official credit rating agencies (Davis and Murphy 2016).

For example, the US P2P lending platform Prosper provides its own credit rating for the investors for credit risk evaluation purposes. The credit rating represents the estimated average annualized loss rate to the investor. This credit rating has 7 levels, from which AA and A represents the lowest risk and HR denotes the highest risk (Prosper 2020; Pokorna and Sponer 2016). Another popular US P2P platform Lending Club assign the loan grade to each loan which is based on the information of the loan application and the FICO score (a credit score provided by third party credit rating agency). The loan grade has 7 levels that are further divided into 5 subgrades. In the Prosper and Lending Club platforms, the interest rates of the loans are determined by the platform based on the assigned credit ratings. Thus, the internal risk evaluation process directly affects the return on investments on both platforms (Lending Club 2020b; Prosper 2020).

The P2P credit scoring systems are not standardized or regulated and as stated by Davis and Murphy (2016), there are also risks in P2P lending platforms acting as financial advisors. This is due to the fact that the P2P lending providers earn their profit as the fees from intermediating the lending process. They typically get the fees when the loan transactions are concluded – no matter whether the borrower will default his loan or not. This can lead to the situation where the platform attempts to maximize the number of issued loans at the expense of the loan quality in the short term. However, in the long term, low default rates serve as a good advertisement for platforms and keeping the default rates low helps the platforms to maintain their reputation (Pokorna and Sponer 2016).

2.5.2 Individual credit risk assessment

Instead of basing the credit investment decisions directly on the platforms' credit scorings, the investors are suggested to do their own investment analysis before lending. However, in P2P lending context, the lack of resources and financial expertise often affects the quality of credit risk evaluation and the tools used for conducting the assessment. As stated by Chen et al. (2014), investors' trust on borrowers affects markedly the lending decisions in P2P lending platforms. Because the trust is difficult to build in the absence of personal contacts and the investors typically do not have enough expertise to assess credit risk exhaustively by themselves, the herding behavior is found typical on P2P lending markets. For example, on the auction-based platforms, it has been found that the investors typically bid on the loan requests that have been bid earlier by other investors (Lee and Lee 2012).

Many P2P platforms provide the historical data of the loan applications for their users. This data can be analyzed and exploited when doing the investment decisions. Both researchers and investors have developed statistical and ML-based credit scoring and default prediction

models to predict the creditworthiness of borrowers and the loan defaults. This thesis focuses on using the ML classification models and FS methods in P2P lending credit risk prediction. The statistical and ML models used in the previous research in P2P credit risk assessment and prediction are discussed in detail in the literature review of this thesis (Chapter 5). In the next chapters, the theoretical aspects of ML-based prediction and FS are introduced.

3 MACHINE LEARNING BASED PREDICTION

ML is a subset of artificial intelligence which has rapidly gained popularity in the recent years due to the increased computing power and explosively grown amount of available data. It can be defined as a field of study which concentrates on exploring and developing the algorithms and statistical models that can independently learn from the data without being explicitly programmed (Liu et al. 2017). One of the most important characteristics of ML algorithms is that they can automatically improve their efficiency during the execution. ML is nowadays playing an important role for example in healthcare, manufacturing industry and image recognition. In the financial field, typical ML applications include for example algorithmic trading and credit scoring (Dietterich 1997a; Michie 1968).

3.1 Different types of machine learning

The simplified taxonomy of ML is represented in Figure 4. The ML can be divided into *supervised* learning and *unsupervised* learning. In the unsupervised learning, the data is unlabeled, and the training data includes only set of input variables without any corresponding target values (Bishop 2006, p.3). A typical example of unsupervised learning is *clustering* in which the unlabeled data is partitioned into groups of similar instances, typically according to some distance measure (Dietterich 1997a). Clustering is frequently used for example in marketing and image analysis.

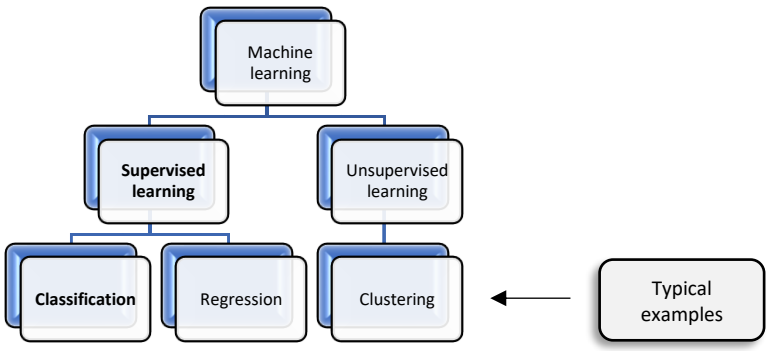


Figure 4. Simplified taxonomy of machine learning.

In contrast to unsupervised learning, in the supervised learning, the ML model is trained with the labeled data. The objective of the supervised learning is to construct a model that explains the target variable in terms of predictor variables (features) (Kotsiantis 2007). A typical example of supervised learning problems is the *classification* problem in which the goal is to assign a discrete category for each instance based on the values of input variables. If the target variable is continuous, the problem is called a *regression* problem (Bishop 2006, p.3). In this thesis, the supervised classification models are used to predict whether the P2P loan is going to be defaulted or not. Default prediction is one of the most commonly examined applications of ML in financial field. As is commonly the case, also in this study, the target variable can take only two values, 1 if the loan is predicted to be defaulted and 0 otherwise. This type of supervised learning task is called a *binary classification problem*. Because this study deals with the classification, in this chapter the focus is on classification techniques.

3.2 Data preprocessing

The raw real-world datasets are rarely in a suitable form for the ML algorithms and they generally need preprocessing before they can be used in the analysis. Typical preprocessing phases in ML include for example handling of missing values, feature encoding and feature scaling.

Handling the missing values is needed because many ML algorithms cannot deal with the missing data (meaning that for some observations, not all variable values are known). The simplest way to handle the missing values is the *complete case analysis* in which the observations with incomplete data are wholly removed from the dataset. If the dataset is large enough and the missing values are considered random (there is no pattern associated with missing values), this technique is suitable (Donders et al. 2006). In cases where the removal of the observations is not a good option, different *imputation techniques* are used to replace the missing data with substituted values. Commonly used imputation techniques include for example *mean imputation* where the missing values are replaced by the mean value of corresponding variable and *regression imputation* in which the missing values are estimated with regression based on the values of other variables (Pelckmans et al. 2005).

Real-world datasets frequently contain both numeric and categorical variables. Because ML models typically accept only numerical inputs, the categorical variables must be converted into numeric form. *Feature encoding* is the process in which the categorical variables are encoded into numerical values. Perhaps the most widely used encoding technique is *One Hot Encoding* (also referred to as *dummy encoding*). In this technique, the categorical feature containing of d classes is transformed into d binary (dummy) variables which indicate the class membership over a corresponding categorical class. The One Hot Encoding is suitable when the classes of

categorical variable have no natural ordering and are not equally spaced. The most significant drawback of One Hot Encoding is that it increases the dimensionality of the dataset because new variables are created for every class of all the categorical variables in the data. In *label encoding* (also referred to as *integer encoding*), an integer is assigned to each class of the categorical variable. The label encoding does not add new columns to the dataset, but the most significant disadvantage of this technique is that it introduces an order for the classes which perhaps does not exist. This can cause problems with some ML models (Potdar et al. 2017).

Datasets also frequently have variables which are measured on very different scales. Many ML algorithms rely heavily on the distance calculations between observations, and particularly in these cases the results can be distorted if the scales of the variables in the dataset vary considerably. To solve this problem, different *standardization* methods have been developed. In the commonly used *min-max normalization* the minimum value of the variable is determined to be 0, the maximum value is set to 1 and all the other values are scaled to lie between 0 and 1. Another popular standardization method is so-called *standard score (z-score) standardization* in which the standardized value is calculated by subtracting the mean value of the feature from the value of the observation and dividing the difference by standard deviation of the feature (Aksoy and Haralick 2001).

3.3 Hyperparameter optimization

Many ML models have optimizable *hyperparameters* that can have a considerable effect on the model's predictive performance. One example of optimized hyperparameters is the minimum leaf size of decision tree algorithm which affects the complexity of the trained tree. The hyperparameters can be optimized manually, following commonly accepted rules of thumb or by automizing the search process. Common automized search techniques include for example the *grid search, random search* and *Bayesian optimization.* In grid search, the prediction model is trained with a user-specified set of values for each of different hyperparameters and the best-performing hyperparameters with regards to some criterion (usually classification error) are chosen (Bergstra and Bengio 2012; Snoek et al. 2012).

Because going through all the combinations of hyperparameters can be computationally very heavy, different techniques have been developed that give sufficient results without testing every possible hyperparameter combination. In the *random search*, the hyperparameter trials are chosen randomly in each iteration from all the possible combinations (Bergstra and Bengio 2012). In the *Bayesian hyperparameter optimization*, the next hyperparameter combination in every iteration is chosen based on the past evaluation results. Bayesian hyperparameter

optimization uses probability model to focus on the range of hyperparameter values that have found to be promising in the previous iterations. Random search and Bayesian optimization can reduce the computational costs considerably compared to grid search, but the drawback of these techniques is that they cannot fully guarantee the optimality of the chosen hyperparameter combination. However, the reliability of the results can be increased by using enough iterations in the search process (Snoek et al. 2012).

3.4 Evaluation of classification models

The classification models are evaluated based on their classification performance, in other words, how well the models can distinguish the instances between the classes under consideration (Japkowich and Shah, 2014, p.12-13). The final evaluation of ML models is a critical phase because different models are typically compared to each other based on the predictive performance (Bradley 1997). In this study, the appropriate evaluation of classification models is essential because the FS methods are validated based on the final classification performance of the classification models.

3.4.1 Confusion matrix

A confusion matrix is a common way to analyze the performance of a classification model on a test set for which the actual values are known. The simplest case of a confusion matrix is the case of binary classification problem, but it can be extended to multiclass problems as well. In the following, the structure and the basic idea of a decision matrix is represented in case of a binary classification problem. In the 2x2 matrix, the instances are divided into four cells regarding their predicted values and actual, observed values (Fawcett 2006). There are four possible alternatives of classes in which the instances can belong to:

1. True positives (TP): instances predicted positive when being actually positive.
2. False positives (FP): instances predicted positive when being actually negative.
3. True negatives (TN): instances predicted negative when being actually negative.
4. False negatives (FN): instances predicted negative when being actually positive.

An example of confusion matrix is represented in Figure 5. The matrix can be used as a basis for calculations of different performance measures. The formulas for some of the most important performance measures are listed in the following (Bradley 1997):

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \quad (1)$$

$$Misclassification\ rate = \frac{FN + FP}{TN + FN + TP + FP} \quad (2)$$

$$Sensitivity = \frac{TP}{FN + TP} \quad (3)$$

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

Accuracy is calculated as the ratio of correctly classified instances to all instances and it is commonly used as an overall performance measure of classification. However, basing the performance evaluation on the accuracy alone has been criticized because it can be misleading especially in the cases where the data is *imbalanced*. This means that considerably more observations belong to one class than to another which is the case in many real-world datasets. For example, in financial distress prediction the proportion of instances that have gone bankruptcy under the period examined is usually markedly lower than the share of non-bankrupt instances. In this case, the accuracy of the classification model that predicts all the observations to the non-bankrupt class would be high even though it classified all the bankrupt instances incorrectly. The *misclassification rate* (also known as error rate) measures the ratio of incorrectly classified instances to all instances. However, this metrics has the same problems than the accuracy as the overall performance measure (Bradley 1997; Powers 2011).

		Predicted		
	N = 100	**Positive**	**Negative**	**Total**
Actual	**Positive**	**TP = 30**	**FN = 20**	**50**
	Negative	**FP = 10**	**TN = 40**	**50**
	Total	**40**	**60**	

Figure 5. Example of a confusion matrix

Because of obvious issues related to the accuracy and misclassification rate as the performance metrics, it is useful to make use of other performance measures as well. For example, *sensitivity* (also referred to as *true positive rate* or *recall*) indicates how often the classifier predicts the actual positive instances correctly. Instead, the *specificity (*also called *true negative rate)* indicates how often the classifier correctly classifies the actual negative instances to negative class (Powers 2011). There are also other performance rates which can be calculated based on the confusion matrix but going through all the measures is not considered necessary in this thesis.

3.4.2 Receiver operating characteristic (ROC) curve

Receiver operating characteristic (ROC) curve analysis is a technique that is used to visualize, organize, and select the classifiers based on their performance (Fawcett 2006). The technique is based on the confusion matrix: in the ROC analysis, the true positive rate is plotted against the false positive rate. The ROC curve offers a graphical presentation of the performance of the classifier. The example of ROC curve is represented in Figure 6.

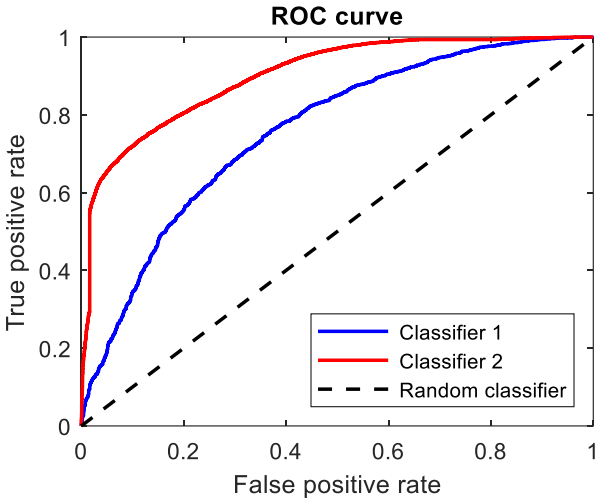


Figure 6. Example of ROC curve

If the classifier perfectly classified all the instances, it would end up in the top left corner of the figure where the true positive rate is 1 and the false positive rate is 0. The worst possible classifier would end up in the bottom right-hand corner where the true positive rate is 0 and the false positive rate is 1. The baseline is frequently drawn to lie on the positive diagonal of the graph which represents the random classifier: it is expected to predict the negative and positive examples at the same rate (Powers 2011). In the example shown in Figure 6, the classifier 2 (red line) outperforms the classifier 1 (blue line) because it is located nearer to the top left corner than the classifier 1. Both classifiers outperform the random classifier (black dotted line).

Based on the ROC curve, the *area under the ROC curve (AUC)* can be calculated which is frequently used as a numerical measure of classifier performance. The AUC has been found to be a more efficient and less biased measure of performance than the overall accuracy and, therefore, it is frequently used as the measure of overall classification performance. AUC value can range from 0 and 1, where 1 implies that the classifier performs perfectly and 0 indicates the worst possible performance.

3.5 Validation of classification results

The classification and model selection results must be validated in a suitable way. The most basic validation method used in ML is called *holdout validation*. It includes splitting the initial data into two separated datasets: *training set* and *test set* (also known as a *holdout set)*. The model is trained with the training data (typically, 70-80% of the data is used to train the model) and the independent test set (typically, 20-30% of the initial data) is used to evaluate the predictive performance of the model on the new, unseen data. The predictive performance on the test set is often referred to as the *out-of-sample performance* (Arlot and Celisse 2010). The use of independent test set is crucial because training and testing with same instances lead to over-optimistic performance estimations. Only a completely independent test set gives a good estimate of the model's performance on new data.

However, separating the test data decreases the sample size of the training data. The split involves so-called bias-variance trade-off: the bigger the training set is, the smaller is the bias of the parameter estimation in training phase and the better is the model accuracy. However, smaller test set leads to the higher variance of the estimate of the test error (Kohavi 1995).

To conduct the model selection and parameter optimization, the initial data is commonly split into three parts: *training set, validation set* and *test set.* The training set is again used to train the model whereas the validation set is used in the model development and model selection phase to estimate the out-of-sample performance. This *validation performance* is used to choose the parameters of the model, to do FS and to conduct any data-driven pre-processing. Finally, the classification performance of the final model is determined using the independent test set which is completely held out of the model selection and development phase. However, holding out both validation and test sets from the model training phase is problematic because the more data is available to use in training the model, the more reliable the results will be (Bishop 2006, pp.23-33).

To answer this problem, k-fold cross-validation (CV) is frequently used in the model development phase. It helps to make use of the whole training data, without risking the independence of the test set. In the CV process, the training set is first split into k equally sized subsamples (folds). After that, the prediction model is trained k times, in each step using k-1 subsamples as the training data and the remaining subsample as the validation data. Typically, k (the number of folds) is set to be 5 or 10 (Kohavi 1995; Mohri et al. 2012, pp.5-6). An example of 5-fold CV is illustrated in Figure 7.

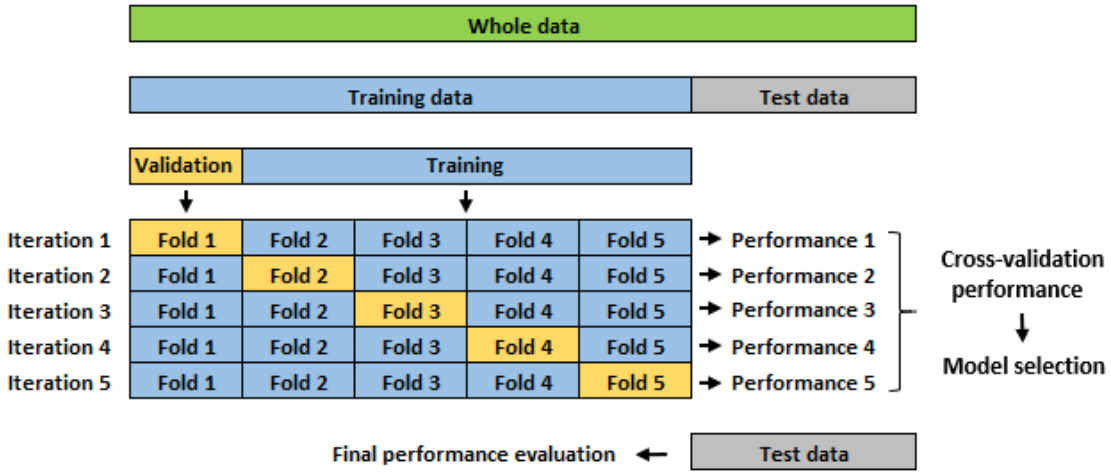


Figure 7. Basic idea of 5-fold cross-validation

The CV performance is determined as the average of model performance over the iterations and the model selection is done based on the CV performance. Finally, the final (out-of-sample) performance is evaluated using the test data (Arlot and Celisse 2010). The procedure for CV described and illustrated above is implemented in model selection phase (for FS and hyperparameter optimization) in this study.

3.6 Classification models of this study

The classification models used in this study are introduced in this chapter. The justification of used models is represented later in Chapter 6.3.

3.6.1 Naive Bayes

Naive Bayes (NB) classifiers are a group of simple supervised classification models that belong to the family of probabilistic classifiers. The basic idea of the models is to estimate the probability that the instance belongs to each class of the target variable and classify the example to the class with the highest probability. The NB classification models are based on Bayes' theorem and rely on a strong assumption of conditional independence between predictors. This means that all the features are assumed to be independent given the value of the target variable, in other words, the predictors should not affect to each other (Provost and Fawcett 2013, p.241; Zhang 2005).

Despite their simplicity and the fact that the conditional independence assumption is rarely fulfilled in real-world applications, the NB classifiers have been found relatively efficient in various classification tasks. It is found that violating the assumption of conditional independence of predictors tend not to hurt the classification accuracy considerably (Provost and Fawcett 2013, p.243). The NB models have been used successfully for example in spam filtering and

text classification and because of their computational lightness, they have also been commonly used for making real time predictions. Furthermore, the NB classifiers are commonly used to benchmark the more sophisticated classifiers in different classification problems (Bishop 2006, p.380; Zhang 2005).

The simplified graphical representation of a NB classification model is shown in Figure 8. The figure illustrates that when conditioned on the class label z, the objects of the predictor vector x (x_1 and x_D) are assumed to be independent. The attributes (predictors) of the model have only one parent node which is the class node (Bishop 2006, p.380; Zhang 2005).

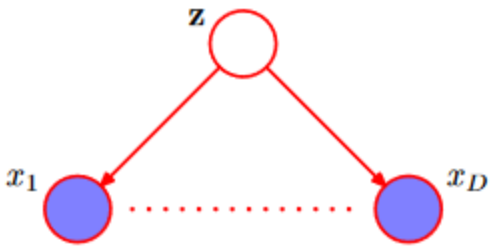


Figure 8. Basic idea of Naïve Bayes classification (Bishop 2006, p.380)

The NB classification models are especially useful when the predictors include both continuous and categorical variables because the features can be represented separately with appropriate models (Bishop 2006, p.381). Two common examples of different NB classifiers are *Gaussian NB* and *multinomial NB* models. The Gaussian NB is usually used with continuous variables and it assumes that the variables follow the normal distribution. The multinomial NB, in turn, is based on the multinomial distribution and is frequently used for classification with discrete variables. (Provost and Fawcett 2013, p.244).

3.6.2 Logistic Regression

Logistic regression (LR) belongs to the family of popular regression analysis methods which are generally applied to explain the relationships between different variables in the data. The LR model can be used in classification problems where the target variable is categorical (or binary as in this study). The basic principle is that the model estimates the probability of class membership of instance over a categorical class which in the context of this study means the probability of belonging to the class of defaulted loans. The model is widely used in the field of ML since it is easy to implement and interpret (Dreiseitl and Ohno-Machado 2002; Mohri et al. 2012, p.129).

The basic principle of the LR is relatively similar to the commonly used linear regression. However, while the model parameters of the linear regression are usually estimated using the

ordinary least-squares estimation, in case of LR the model parameters are determined using maximum likelihood estimation. This means that the parameters are chosen so that they are the most likely values of the model parameters in the used data, in other words, these parameters optimize the value of the used likelihood function (Bishop 2006, pp.205-206; Dreiseitl and Ohno-Machado 2002).

In contrast to linear regression, the LR attempts to estimate the class membership probability of an instance which is done using the logarithmic transformation of *odds ratios* (also known as *log-odds*) of the independent variable. The odds are determined by dividing the probability of an instance belonging to the certain class by the probability of an instance not belonging to that class. The basic notation of the LR model with two explanatory variables can be represented as follows (Hosmer et al. 2013, p.7; Peng et al. 2002):

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}} \qquad (5)$$

In the equation, $ln\left(\frac{p}{1-p}\right)$ is the log-odds of the independent variable, p is the probability that an instance belongs to the certain class, x_1 and x_2 are the explanatory variables, β_0 is the constant term, and β_1 and β_2 are the estimated regression coefficients. The class membership probability (p) can be calculated from the basic equation by solving the equation for p as illustrated above. In the last represented equation, the constant e denotes the Euler's number and the other symbols are the same as described earlier (Hosmer et al. 2013, p.7; Peng et al. 2002). The class membership probabilities are used to classify the instances into different classes of a categorical target variable.

3.6.3 Decision Tree

Decision tree (DT) model is a simple and very commonly used prediction model which is developed in the form of a tree-like structure. It has become very popular because it is quick to train and easy to interpret. The DT models can be conducted on both categorical and continuous data, and they have been found to be relatively efficient in the predictions (Mohri et al. 2012, p.194). However, one significant drawback is that DTs are tending to overfit the data they are trained with (Kotsiantis 2007). The basic principle of the DT classification algorithm is to classify the population into branch-like segments, constructing an inverted tree which has three types of nodes: root node, internal nodes and leaf nodes (Song and Lu 2015).

The root node is the topmost node of the tree, having two or more branches and representing some test or rule which is used to do the decision that will result in the division of the data into two or more subsets. The internal nodes are used to further divide the data in smaller subsets

according to some defined rules. Finally, the leaf nodes represent the final result of the classification and indicate the value of the target variable (Song and Lu 2015). In the DT model, the instances are classified beginning from the topmost (root) node of the tree, moving downwards until a leaf node (Safavian and Landgrebe 1991).

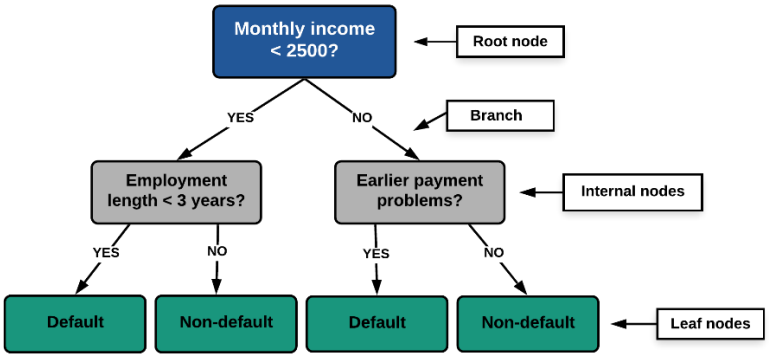


Figure 9. Example of a binary decision tree

Figure 9 shows the typical example of a binary DT. The tree represents the binary classification model which is used to predict whether the borrower is going to default his loan or not. The root node represents the most influential feature in classifying the defaults (monthly income). The internal nodes define the other rules (according to the employment length and earlier payment problems) which are used in classification. The leaf nodes represent the final class labels (whether the loan is defaulted or not). Each parent node is connected to the child node (or leaf node) with a branch which shows the outcome of the test (Song and Lu 2015).

3.6.4 Random forest

Despite their obvious advantages, the DT classifiers have also some significant drawbacks such as instability, overfitting and relative inefficiency compared to more sophisticated classification models in many real-world cases (Kotsiantis 2007; Malekipirbazari and Aksakalli 2015). Therefore, different *ensemble methods* have been developed to enhance the predictive performance of DTs. In ensemble methods, multiple classifiers are used to solve the same problem. In *bagging*, also referred to as *bootstrap aggregation*, multiple decision trees are built by uniformly constructing several bootstrap replicates from the initial training set by using sampling with replacement. The results of individual trees are averaged to make more reliable predictions. Another ensemble method called *boosting* typically uses iterative learning to assign weights for the instances in the training data. Misclassified instances get bigger weights than the correctly classified ones, and therefore the final classification model will focus more on the difficult instances. This often leads to the improved classification performance (Dietterich 2000).

Random forest (RF) is a bagging ensemble algorithm which constructs a combination of tree-based predictors to predict the outcome. The RF model can be used in both classification and regression tasks. In the RF classification model, multiple decision trees are built by selecting the input features of each tree randomly from the set of all available features. The individual trees are trained with a random feature subset and the output is predicted based on this information. The final output of the model is the output class with most votes from individual trees (Breiman 2001).

Because the final outcome of the model is averaged over multiple trees, the RF method helps to reduce the variance of the result compared to a single DT classifier. Besides that, randomizing the selection of input features de-correlate the individual trees in the forest which is an advantage over the traditional bagged trees (Breiman 2001; Malekipirbazari and Aksakalli 2015). The RF model as a bagging method makes use of randomly chosen bootstrap replicates of the original training data to train the individual trees in the forest, which leaves some observations out of the training phase in case of each tree. These observations are called *out-of-bag* observations, and they can be used for example in validating the RF model and in estimating the feature importance (see Chapter 4.6.4 for details).

4 FEATURE SELECTION

During the latest decades, the quantity of available data has exploded. At the same time, more sophisticated ML techniques have been developed to address more complex tasks (Chandrashekar and Sahin 2014). For example, in the gene selection and text classification problems, the number of features can be extremely large (Guyon and Elisseef 2003). In the financial field, for instance the datasets used for financial distress prediction and predicting the financial crises are typically very complex and high-dimensional (Liang et al. 2015).

Different dimensionality reduction techniques have been introduced to address the problems of high dimensionality of datasets (Chandrashekar and Sahin 2014). The dimensionality reduction methods are commonly divided into two groups: *feature selection* and *feature extraction* methods. Feature extraction is used to reduce the dimensionality of the data using transformations and derived values of the original features to generate the smaller subset of features that still contains the most discriminatory and relevant information of the original variables (Khalid et al. 2014). In this study, the focus is on the FS methods and therefore, the feature extraction methods are not introduced in more detail.

FS, instead, can be defined as a process in which a subset of relevant features is chosen from the group of all existing features (Dash and Liu 1997). In the process, the irrelevant or

redundant features are removed and only the most relevant features are left in the final model (Bolon-Canedo et al. 2013). The *relevance* of a feature has multiple definitions and can be defined differently from varying perspectives. Frequently, the relevance of the feature is measured by calculating some dependency measure (typically correlation) between target variable and the feature (Hall 1999).

However, because this thesis focuses on FS for a practical classification application, it is reasonable to define the relevance with respect to the classification accuracy (or more widely, classification performance). Blum and Langley (1997) proposed a definition for relevance based on *incremental usefulness.* According to this definition, a feature *f* is relevant (incrementally useful) if including this feature in a feature set *A* leads to better accuracy using the learning algorithm *L* than the feature set *A* without adding the feature *f*. The other way around, the feature *f* is relevant if it cannot be removed from the feature set *A* without a deterioration of classification accuracy obtained using learning algorithm *L* (Blum and Langley 1997; Kohavi and John 1997).

In contrast to the relevance, the *redundancy* refers to the inter-relationships between different features in a feature set. The redundancy of the feature is typically defined in terms of feature correlation but other dependency measures such as mutual information can be used in definition as well. It is commonly accepted that the feature is redundant if it is perfectly correlated with another feature or can perfectly be determined by using a linear combination of other features in the feature set. However, the correlation does not always need to be perfect but also strongly correlated features can be considered redundant (Yu and Liu 2004).

As proposed by Dash and Liu (1997), typical FS process can be represented as a cycle where the process is split into four key steps (shown in Figure 10).

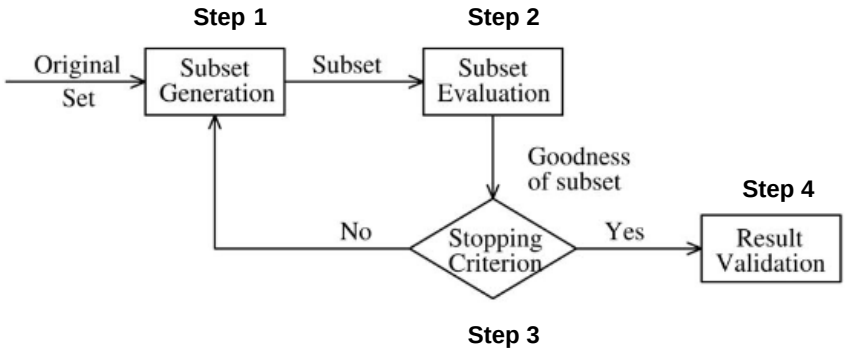


Figure 10. Key steps of feature selection process (Dash and Liu 1997)

In the first step of the process, a candidate feature subset is generated from the original feature set using some search strategy. After that, in the second step, the generated candidate subset

is evaluated, and its goodness is compared to the previous best subset using some evaluation criterion. If the new candidate subset is better, it replaces the previous best subset. The steps 1 and 2 are repeated until some predetermined stopping criterion is met (step 3). The stopping criterion can be met, for example, if addition of any feature to the previous subset does not lead to a better performing subset of features (Dash and Liu 1997; Liu and Yu 2005).

After the stopping criterion is met, the final subset of features is validated (step 4). In fact, the validation step is not a part of the actual FS process, but it is an essential part of using any FS method. The validation can be done, for instance, using the previous knowledge of the best optimal feature subset or by comparing the results with competing FS methods using artificial or real-world datasets (Dash and Liu 1997; Liu and Yu 2005).

4.1 Different types of feature selection

As illustrated in Figure 11, the FS can be split further into different groups. Based on the label information, FS is commonly divided into three classes: *supervised, semi-supervised and unsupervised* FS (Chandrashekar and Sahin 2014). The *unsupervised* FS is applied when no class labels are known but one still needs to select a subset of most significant features with regards to a pre-defined criterion, for example, variance or correlation. The unsupervised FS is commonly used in clustering which is a popular example of unsupervised learning (Guyon and Elisseef 2003).

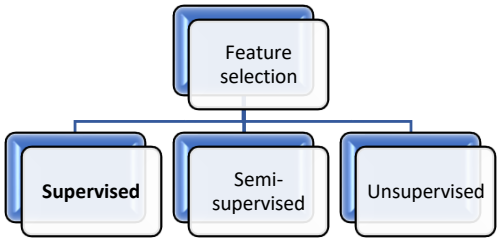


Figure 11. Different types of feature selection

In the cases where only some of the class labels are known, the *semi-supervised* FS can be used to exploit both labeled and unlabeled data to evaluate the relevance of the features and to select the optimal feature subset. The semi-supervised methods are usually applied in the cases where the labeled real-world examples are difficult to find, for example, in the areas of medical diagnosis and fraud detection (Sheikhpour et al. 2017).

In the *supervised* FS, the class labels for all observations are known beforehand and this information is used in the selection process of the optimal feature subset (Saeys et al. 2007). The supervised FS methods are applied in the context of supervised learning and they can be

used with both classification and regression problems. This thesis focuses on the supervised binary classification and, therefore, the supervised FS is under consideration in this study.

4.2 Main classes of feature selection methods

As illustrated in Figure 12, the FS methods are commonly divided into 3 main classes based on how these methods combine the FS process with the construction of classifiers. These main classes are *filter* methods, *wrapper* methods and *embedded* methods (Blum and Langley 1997).

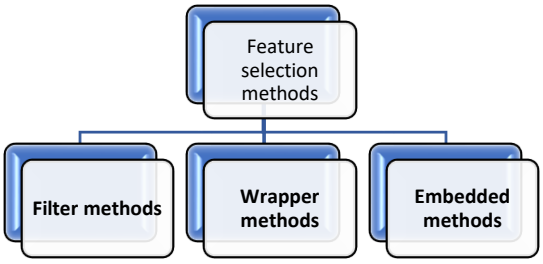


Figure 12. Taxonomy of feature selection methods

The basic principle of the *filter* methods is to rank and order the features in the data by some evaluation criterion and to include the most relevant features in the final feature subset. In these methods, the features are given a ranking score which indicates the relevance of the feature to the output class (Chandrashekar and Sahin 2014). After the ranking, a user-specified number of features with the highest ranks are used in the classification. Alternatively, a minimum threshold for the value of an evaluation criterion can be assigned to remove features with the score below the threshold (Saeys et al. 2007).

The basic idea of filter-based FS is represented in Figure 13. As it can be seen from the figure, the filter methods are used to filter out the irrelevant features as a part of the pre-processing phase and therefore before the actual classification algorithms are applied. This makes them independent of the classification algorithm, and therefore the filter methods can be used with any algorithm (Blum and Langley 1997). In addition, the filter methods are typically computationally light, fast and found to be efficient with many datasets (Liu and Yu 2005). Typical examples of filter-based FS techniques are correlation-based FS and ReliefF algorithm.

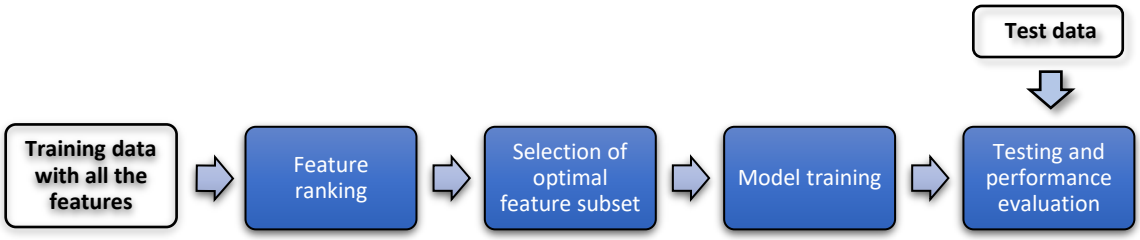


Figure 13. The basic idea of filter-based feature selection

However, it is possible that the feature subset obtained using a filter method is not optimal because the filter methods do not consider the inter-correlations of features. This can lead to the selection of a redundant subset. The filter methods also commonly fail to take into account the feature interactions and there is also no universal method for choosing the number of features for the final subset (Chandrashekar and Sahin 2014; Saeys et al. 2007).

In contrast to the filter methods, the *wrapper* methods are not applied in the pre-processing phase and they deploy the specific classification algorithm in the FS process (Liu and Yu 2005). The basic idea of wrapper-based FS is illustrated in Figure 14. The wrapper methods generate multiple subsets of features and evaluate their performance usually by using some classification performance metric as the evaluation criterion. The specific classification model is run multiple times with different (smaller) combinations of features from the original feature set and the feature subset with the highest evaluation (typically with the highest classification accuracy or the smallest classification error) is selected to be the final feature subset. After that, the final classification is conducted with the chosen subset of features, and the performance of the model is evaluated (Kohavi and John 1997). Common examples of the wrapper methods are sequential FS methods (see Chapter 4.6 for more details).

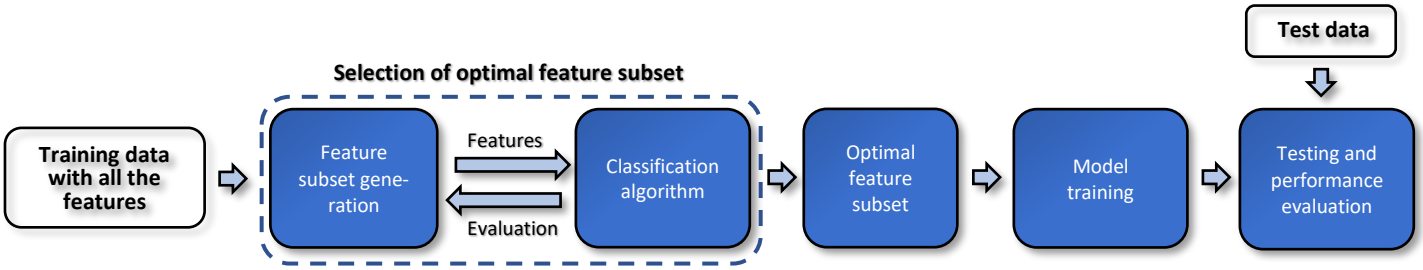


Figure 14. The basic idea of wrapper-based feature selection

The advantages of the wrapper methods include the ability to consider the inter-dependencies of features and usually obtained high classification accuracy. The wrapper methods do not either require the user-specified thresholding for the evaluation criterion or for the number of features. However, a major drawback of the wrapper methods is that they are computationally very intensive because of the large number of computations required to build the optimal feature subset. In wrapper-type FS methods, the classifier is trained, and the classification accuracy of the model is calculated in each iteration of the algorithm using a different feature subset (Kohavi and John 1997). This is more computationally heavy than running a filter-type feature ranking method once with a frequently less complicated evaluation criterion. Also, the proposed feature subset always depends on the used classification algorithm and therefore it cannot be generalized directly to other classification models (Chandrashekar and Sahin 2014). Furthermore, the wrapper methods are found to be more prone to overfitting than the filter

methods (Saeys et al. 2007). To reduce the overfitting problem, different CV performance measures are frequently used in the evaluation of goodness of different feature subsets (Kohavi and John 1997).

The main idea of the *embedded* methods is to include the FS search into the model training phase. The most significant advantage of these models is that they can decrease the computational costs compared to the wrapper methods, still including the interaction with the classification model (Lal et al. 2006). However, as in the case of wrapper methods, the optimal feature subset proposed by embedded FS methods is again dependent on the specific classifier (Lal et al. 2006). Some embedded methods do not take the interactions between features into account either (Saeys et al. 2007). The basic idea of embedded FS is illustrated in Figure 15.

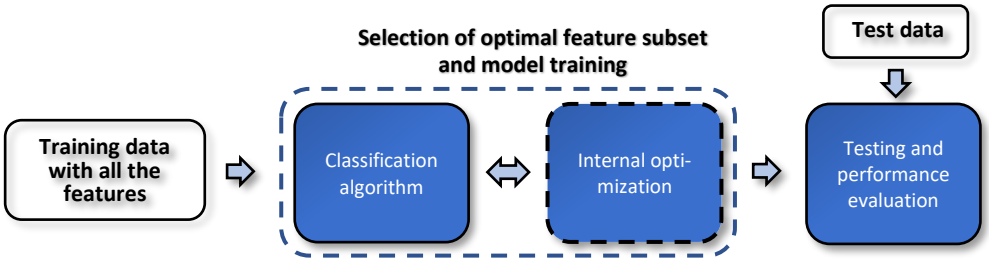


Figure 15. The basic idea of embedded feature selection

As the figure shows, FS and model training phases cannot be separated in case of embedded methods. The embedded FS methods include a great variety of different techniques and therefore they are more difficult to analyze as a group than the filter and wrapper methods (Lal et al. 2006). *Built-in* embedded methods exploit the built-in FS processes of the ML models. A typical example of built-in embedded FS is the RF-based FS which exploits the predictor importance score of RF algorithm (see Chapter 4.6.4 for closer details).

Pruning based FS methods start with the full set of features and remove features by forcing the corresponding feature coefficients to zero. An example of this type of embedded methods is the support vector machine recursive feature elimination (SVM-RFE). In the *regularization* methods, different weights are set to the feature coefficients that indicate the importance of each feature in the model. These methods are commonly used in ML models to assist in reducing the overfitting and underfitting problems. A common example of this type of methods is LASSO (least absolute shrinkage and selection operator) regularization technique (Azim and Ahmed 2018, pp.38-39; Lal et al. 2006).

4.3 Search strategies

The search strategy (or generation procedure) plays an important role in FS process. It is used to generate candidate feature subsets from the full set of existing features (Liu and Yu 2005). In the ideal case, the optimal feature subset would be chosen by trying all the possible combinations of the original feature set and comparing their goodness with regards to some evaluation measure. This kind of search strategy is called *exhaustive* search (Dash and Liu 2003).

However, the number of candidate subsets rises exponentially when the number of features increases and for the dataset including N features, the number of possible feature combinations becomes 2^N. Going through all these combinations would be computationally very intensive, especially for high-dimensional data, and therefore, different search strategies have been developed to guide the search. According to Dash and Liu (1997), the search strategies are divided into three types: *complete, heuristic* and *random* search.

In the *complete* search, the optimality of the found best-performing feature subset is secured according to the used evaluation criterion. This search strategy is used to conduct a complete search with regards to the used evaluation function (Dash and Liu 1997). One commonly used complete search algorithm is a *Branch & Bound* algorithm which guarantees the optimality of the feature subset by backtracking (Dash and Liu 2003). The *heuristic search* algorithms are typically fast, but unlike the complete search algorithms, they typically do not guarantee the optimality of the selected subset. These algorithms generate and evaluate multiple feature subsets to optimize a chosen objective function. The feature subsets are created typically by searching around the search-space or by generating different solutions to an optimization problem. The heuristic search algorithms are commonly used in FS because they have found to produce sufficient approximate solutions with relatively low requirements of time and memory space (Chandrashekar and Sahin 2014). Typical examples of FS methods using heuristic search are *ReliefF* algorithm and *Fisher score* method (Freeman et al. 2015).

In the *random* search, the search process is started by selecting randomly an initial subset of features. After that, the process is continued either by introducing randomness into the traditional sequential search approaches or by conducting the subset generation process based on some probability distribution (Liu and Yu 2005). One common example of random search algorithms is the *random-start hill-climbing* technique (Dash and Liu 2003; Liu and Yu 2005).

4.4 Evaluation criteria

Another important factor of FS process is the evaluation criterion used to evaluate the generated subset of features. It is noteworthy that the choice of the evaluation criterion is essential

because the subset which is optimal with regards to some evaluation criterion is not guaranteed to be optimal with regards to other criteria. Dash and Liu (1997;2003) divide the measures into 5 categories which are introduced in the following.

The *distance* measures (also referred to as divergence, separability or discrimination measures) are frequently used in FS algorithms. The logic of these measures can be demonstrated in the context of a two-class problem. Because the aim is to find the features that distinguish the two classes as perfectly as possible, the feature F_1 is preferred to another feature F_2, if the feature F_1 results in a larger spread between the two-class conditional probabilities than feature F_2. In other words, the feature subset is considered optimal when it maximizes the class separability. A typical example of distance measures is the *Euclidean distance* (Liu and Yu 2005; Dash and Liu 2003).

In contrast to the distance measures, *consistency* measures do not try to maximize the class separability. Instead, the aim is to maintain the discriminating power of the data obtained by using the original set of features while decreasing the number of features. Thus, using this measure, the aim of the FS is to find the smallest consistent feature subset which can distinguish the classes as accurately as the full set of features (Dash and Liu 2003; Almuallim and Dietterich 1994). The feature subset is consistent if no two instances having the same feature values have a different class membership. Because the pure consistency can be impossible to reach in real-world datasets, the user-defined inconsistency rate is frequently used and the minimum feature subset with regards to this constraint is considered as optimal (Dash and Liu 1997).

Dependency measures (also referred to as correlation or similarity measures) are used to measure the possibility to predict the value of one variable with other variables in the dataset. In the context of classification, the correlation measures are frequently used to measure how strongly the features are correlated with the class labels. The features that are in a strong relationship with class labels are preferred to the ones with low correlation (Liu and Yu 2005). Furthermore, dependency measures can be used for identifying the redundant variables. Based on the correlation, the features which have high correlation with other features (predictors), can be removed to reduce the problems of redundancy of the dataset (Dash and Liu 1997).

Information measures typically measure the information gain of the feature. In the FS based on the information measure, the feature F_1 is preferred to feature F_2 if the information gain obtained by including the feature F_1 in the feature subset is greater than the information gain obtained by including the feature F_2 (Liu and Yu 2005). The information gain obtained by

adding a feature into the feature subset can be measured by the difference between prior uncertainty (before the feature was included) and posterior uncertainty (after the feature has been included) using some measure of uncertainty. A commonly used measure of uncertainty is entropy (Dash and Liu 1997).

The *classifier error rate* measures (also known as accuracy measures) are typically used in case of wrapper FS methods. The classifier error rate is in a relation with the classification accuracy because it can be derived by subtracting the classification accuracy from unity (Kohavi and John 1997). As the name indicates, the classifier is included in the process of evaluating and selecting the feature subset. Obviously, the feature subset with the smallest classifier error rate (and highest classification accuracy) is selected to final classification (Liu and Yu 2005).

4.5 Validation of feature selection methods

The final part of the process of using any FS method is validation. As proposed by Liu and Yu (2005), the simplest way to validate the results is to use the previous knowledge about the data. If the optimal subset of features is known beforehand, the performance of a FS method can be evaluated based on this information. However, in case of this study (and many real-world applications), the information about the optimal subset of features is not available, and other (indirect) validation approaches need to be used.

To validate the results of a FS method, it is possible to compare the results with competing FS methods with regards to some evaluation measure (Dash and Liu 1997; Liu and Yu 2005). Typically, the final classification accuracy is used for this purpose. The validation can also be done by conducting a simple “before-and-after” experiment. In this approach, the classification error rate obtained using the original model (before FS) is compared to the error of the model in which the optimal feature subset is used (after FS) (Liu and Yu 2005).

4.6 Feature selection methods of this study

The selected FS methods used in this thesis are introduced in this chapter. The justification of each selected method is represented in Chapter 6.3.

4.6.1 Maximum-relevance-minimum-redundancy feature selection

The *maximum-relevance-minimum-redundancy* (MRMR) FS approach is a filter-type FS method that was first introduced by Peng et al. (2005). It is based on the idea of two popular FS approaches: *maximal relevance* FS and *minimum redundancy* FS. In the maximal

relevance FS, the features that have the highest relevance to the target class according to some measure (typically correlation or mutual information) are chosen into the final feature subset that is used for classification.

However, it is found that a feature set including the features with the highest relevance to the target class does not always lead to the best classification performance because high redundancy between some of these features can distort the results of classification. Therefore, the minimum redundancy FS is often used in combination with the maximal relevance FS to select the features that have maximal relevance with the target class while minimizing the redundancy between features (Ding and Peng 2003; Peng et al. 2005). This kind of approach is called maximum-relevance-minimum-redundancy FS.

In many MRMR algorithms (including Matlab's MRMR algorithm which is used in this study), the *mutual information* (MI) is used as a measure of dependence between variables. MI defines a quantity of information that can be obtained from one random variable by having knowledge of another. In other words, it measures the reduction in uncertainty of one random variable if the knowledge about another random variable is available (Peng et al. 2005). The MI of two discrete variables (*x* and *z*) can be calculated based on their joint probabilistic distribution $p(x,z)$ and the corresponding marginal probabilities $p(x)$ and $p(z)$ using the following formula (Ding and Peng 2003):

$$MI(x,z) = \sum_{i,j} p(x_i,z_j) \log \frac{p(x_i,z_j)}{p(x_i)p(z_j)} \qquad (6)$$

In MI-based MRMR algorithms, the relevance of features is typically measured by calculating the MI of target variable and the corresponding feature whereas the redundancy between features is measured by calculating the pairwise MI of different features. The relevance of a feature set S with respect to a target variable y (V_S) and the redundancy of the same feature set S (W_S) can be represented as follows (Ding and Peng 2003):

$$V_S = \frac{1}{|S|} \sum_{x \in S} MI(x,y) \qquad (7)$$

$$W_S = \frac{1}{|S|^2} \sum_{x,z \in S} MI(x,z) \qquad (8)$$

In the first formula, $MI(x,y)$ denotes the MI of target variable and feature *x* calculated as represented in formula 6. In the second formula, $MI(x,z)$ represents the pairwise MI between features *x* and *z*. In MRMR FS, the feature set that maximizes the relevance while minimizing the redundancy is chosen for classification (Ding and Peng 2003). MI-based FS algorithms are

widely used for FS with discrete features and in cases where there are mixed types of features (both continuous and discrete features) in the data, the continuous features are frequently discretized by binning (Sharmin, Ali, Khan, Shoyaib 2017).

4.6.2 Chi-Square feature selection

Chi-Square FS method is a filter-type FS technique which relies on Chi-Square test of independence. Chi-Square test is a simple statistical test belonging to the family of univariate analysis which is usually used to examine the association (dependency) between two categorical variables. The Chi-Square FS method can be seen as an alternative for popular correlation-based FS method in cases where the feature set includes categorical variables because the correlation is not a meaningful dependency measure in these cases. Therefore, another metrics is needed to measure the relationships between predictors and the response variable (Zheng, Wu, Srihari 2004).

In Chi-Square test, a test statistic is calculated which can be used to investigate the magnitude of the dependence. The value can be compared to the critical value of Chi-Square distribution to investigate the statistical significance. In FS context, assuming that the variables examined are the target variable (variable 1) and a chosen feature from the feature set (variable 2), the Chi-Square test statistic can be determined using the following formula (McHugh 2013):

$$X^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \qquad (9)$$

In the formula, $O_{r,c}$ denotes the observed frequency at level r of variable 1 (the target variable) and at level c of variable 2 (the feature examined). $E_{r,c}$, in turn, denotes the expected frequency at level r of variable 1 (the target variable) and at level c of the variable 2 (the feature examined) (McHugh 2013). In Chi-Square based FS, the relationships between features and the target variable are investigated individually using multiple Chi-Square tests. Then, the variables are ranked based on the test results, and the features with the strongest association with the target variable are used in final classification (Zheng, Wu, Srihari 2004).

Advantages of Chi-Square FS include its simplicity, intuitiveness and computational lightness. However, the most significant drawbacks of the method are that it fails to avoid the redundancy of the chosen feature subset and to consider the inter-relations between features. Because the Chi-Square FS method is well suited for categorical variables but is not directly applicable for continuous variables, continuous variables are frequently discretized by binning to handle the data with mixed types of features.

4.6.3 Sequential forward selection

Sequential forward selection (SFS) algorithm is an iterative wrapper-type FS method which begins with an empty feature set and in each step of the process, adds a feature to the subset which improves the value of the chosen objective function (evaluation criterion) the most. The used evaluation criterion is typically related to the measure of classification performance, typically the classification accuracy or classification error. In the first step of the algorithm, each feature is added individually to the empty feature set and the value of the objective function is calculated using the chosen classification model. Then, the feature which provides the best value of the evaluation criterion is added to the empty feature set. In the second step, the feature that performs best in a pair with the feature selected in the first step (in terms of the chosen evaluation criterion) is included in the set of selected features (Chandrashekar and Sahin 2014).

In the following steps, all the features that have not yet been included in the subset of chosen features are considered for selection, and the feature whose inclusion results in the biggest improvement in the value of the objective function is added to the feature subset. The process is typically proceeded until a pre-determined stopping criterion is altered, for instance, until the classification accuracy is not increased anymore by adding any new feature to the subset (Chandrashekar and Sahin 2014). Alternatively, as often done for comparison purposes, the search can be accepted until all the features have been included in the feature set.

The advantage of SFS is that as the wrapper method, it typically leads to high classification performance. This can be seen logical because the measure of classification performance is frequently used as the evaluation criterion in subset evaluation. In addition, the SFS algorithm takes into account the interaction with classifier. However, SFS has also some major drawbacks, including computational expensiveness and the tendency to get stuck in local optima of the objective function. SFS has also so-called *nesting* property which means that once the feature has been added to the subset of chosen features, it cannot be removed in following steps even though the removal would lead to an improvement in objective function value. This can lead to the selection of non-optimal feature subset. As other wrapper methods, the SFS method is also relatively prone to overfitting problems (Reunanen 2003).

4.6.4 Learning-model based feature ranking

An embedded learning-model based feature ranking (referred later to as LMBFR) FS method is also used for FS in this study with DT and RF classifiers. The method has been widely used for FS in previous research because of its intuitiveness and efficiency (Jin and Zhu 2015; Xia et al. 2017; Zhou et al. 2019). The basic idea of the method is simple to understand: first, the

corresponding classification model is trained and the estimates of relative importance of features in the model are estimated. Then, the features with the highest estimated feature importance are used as the final feature subset. The proposed feature subset is classifier-specific, and it does not necessarily generalize to other models but usually provides good classification accuracy with a corresponding classifier (Ishwaran 2007; Kazemitabar et al. 2017).

There are different ways to compute the relative importance of features in the classification models. In case of DT model, the feature importance is frequently calculated by summing up the weighted impurity reductions for all nodes where the corresponding variable is used (Kazemitabar et al. 2017). This measure is called *mean decrease impurity importance* (MDI) and it can be represented for variable X_m using following formula (Louppe et al. 2013):

$$MDI(X_m) = \sum_{t \in v(s_t) = X_m} p(t)\Delta i(s_t,t) \qquad (10)$$

In the formula, t denotes a node of the tree, s_t denotes a certain split, $p(t)$ denotes the proportion of samples reaching node t, and $\Delta i(s_t,t)$ represents the reduction in impurity when the variable $v(s_t)$ is used in split s_t. Thus, the whole formula represents the sum of impurity reductions over all the nodes where a split was made on the feature examined, while weighting the impurity reductions to account for the size of the node (Kazemitabar et al. 2017; Louppe et al. 2013).

For the ensemble of DTs such as RF, the above-mentioned importance measure can be averaged over the trees. Alternatively, the feature importance in case of RF can be estimated using the *mean decrease accuracy* (MDA) measure. This metric measures how strongly permuting randomly the out-of-bag samples of a corresponding feature affects the model accuracy and is also referred to as *permutation importance (*Breiman 2001; Louppe et al. 2013). The permutation importance can be calculated for feature X_j with the following formula (Louppe et al. 2013):

$$MDA(X_j) = \sum_{t=1}^{T} \frac{1}{T} A(D_{O,t}) - A(D_{P,t}) \qquad (11)$$

In the formula, $D_{O,t}$ denotes the original set of out-of-bag samples in case of tree t and $D_{P,t}$ denotes the set of out-of-bag samples after permuting randomly the samples of variable X_j. T denotes the number of trees, $A(D_{O,t})$ represents the out-of-bag classification accuracy before the permutation and $A(D_{P,t})$ represents the accuracy after the permutation. Therefore, the permutation importance of a variable is calculated by summing up the differences between the out-of-bag accuracies before and after the permutation in case of each tree and averaging this

sum over the trees in the forest. Because in RF model the feature set is chosen randomly for each tree (and all the features are not used in all trees of the forest), the importance score is assigned to be 0 in cases where the corresponding feature is not used to construct the corresponding tree (Louppe et al. 2013).

5 LITERATURE REVIEW

In this chapter, the literature review of the research topic is conducted to gain an understanding of previous research and the methods used in them. According to Webster and Watson (2002) the literature review eases the process of theory development, introduces the areas where the comprehensive research has already been conducted and helps to identify the research gaps in the previous research. As stated by Hart (1998, p.13-14), the literature review can also be used as a tool in justifying the research topic and methodology. It also helps to narrow the scope of the research and to identify the major issues and debates about the research topic. Also, the literature review helps in formulating and focusing the research questions and hypotheses. In the following subchapters, the terminology of credit risk management and assessment, the applied literature review method and the literature search process are first described. Then, the previous research related to the research topic is reviewed.

5.1 Definitions

To clarify the content of this literature review, it is essential to define some terms generally used in the credit risk research. Bielecki and Rutkowski (2004, p.11) define the *credit risk* as the risk which is associated with any credit-linked event such as the realized changes in credit ratings, variations in a credit spread or the event of default. In contrast, they define the *default risk* as the possibility that a counterparty of financial contract cannot meet the commitments stated in the contract (Bielecki and Rutkowski (2004, p.11). Therefore, by definition, the default risk as a term is more narrow than the credit risk – default risk focuses exclusively on the probability that default event occurs whereas the credit risk takes into account any change in the credit quality (even the improvement in borrower's creditworthiness). It is also noteworthy that these two terms are often used interchangeably even though, as discussed, they do not necessarily have exactly the same meaning.

Credit risk management, in turn, is a process that involves identification and analysis of the risk factors of credit approval. It includes measurement, prediction and control of the risks associated with the credit activities (Abdou and Pointon 2011; Van Gestel and Baesens 2008, p.3). According to Crook et al. (2007), *credit risk assessment* is an essential part of credit risk management and can be defined as a process that utilizes the risk assessment tools to

manage the credit account of the borrower. The process starts from pre-screening a potential application and proceeds as continued monitoring through the lifetime of the account and possible write-off. Traditionally, the credit risk assessment has been one of the most important tasks of banks. The standardized and strictly verified financial information of credit applicants has been collected by banks and this information has been used to assess the credit risk of individual borrowers (Klafft 2008; Pokorna and Sponer 2016).

According to Abdou and Pointon (2011), two ways have been commonly used to do the lending decision: *judgemental evaluation* and *credit scoring*. The *judgemental evaluation* technique is the most basic way to determine the creditworthiness of the borrower. In this technique, each loan application will be evaluated individually by the experienced decision-maker. The results depend heavily on the common sense and experience of the credit analyst and can be distorted by subjective prejudices and individual preferences. However, there are also some advantages of using judgemental evaluation technique such as taking into account the qualitative aspects and exploiting the credit analyst's prior knowledge (Abdou and Pointon 2011).

Because of the obvious drawbacks of the evaluation technique based on the credit analyst's subjective assessment, the more sophisticated, quantitative ways of credit risk assessment have been developed (Abdou and Pointon 2011). In the *credit scoring* models, the information of the loan application is analyzed with statistical methods to set a numerical credit score for each applicant. This credit score is then used as the decision support tool when doing the final lending decision (Onay and Ozturk 2018). The advantages of credit scoring techniques compared to judgemental techniques include for example time and cost savings and more consistent and objective results of the analysis (Abdou and Pointon 2011). For these reasons, credit scoring has become an important part of financial institutions' credit risk management.

The information that is exploited to assess the creditworthiness of the borrower can be divided into two types based on the quality of the information: *hard information* and *soft information*. The traditional credit scoring models rely mostly on hard information which is quantitative, easy to collect and store and can be analyzed with relatively small effort on pre-processing (Liberti and Petersen 2018). Examples of hard information used in credit scoring models are for example age, income level and debt to income ratio of the borrower (Zhang et al. 2016). In contrast, the soft information often needs pre-processing to be analyzed with statistical methods and is more reliable on interpretation and opinions of the observer (Liberti and Petersen 2018). Examples of soft information that has been used in credit scoring area include an attitude of the loan description texts, the characteristics of borrower's image and the information obtained from the borrower's profile in social media (Dorfleitner et al. 2016; Zhang et al. 2016).

The conventional credit scoring systems typically aim to estimate the default probability and minimize the misclassification error in discriminating the attractive borrowers from the ones who are more likely to default. This approach has often been criticized especially in case of P2P lending because it does not guarantee the maximum profit for the lender. Defaulted loans can sometimes be partially recovered, and the higher interest rates associated with the loans with higher risk of default can sometimes even fully compensate the costs associated with the defaults of the loans (Serrano-Cinca and Gutierrez-Nieto 2016). Therefore, different models have been developed for P2P loan evaluation purposes which aim to measure the expected profitability of loans instead of focusing on forecasting the default probability. These methods are referred to as *profit scoring* models (Ye et al. 2018). In this review, the focus is on the credit scoring models but some approaches for profit scoring are also described.

5.2 Methodology

The literature search method is in a significant role in ensuring the quality of the literature review. In this study, the search process of the most significant literature is done based on the three-step systematic selection approach proposed by Webster and Watson (2002) and can be represented as follows:

1. The process is started by searching for the major contributions to the research topic in the leading journals, also considering the relevant articles outside the primary discipline.
2. In the second step, the process is continued by *going backward* to identify prior relevant articles. This is done by checking the citations in the articles found in the first step.
3. In the third step, the process proceeds by *going forward* to find the relevant articles citing to the research found in the previous steps of the process.

In each step, the found relevant articles are added to the set of reviewed articles. Finally, the process is finished when there are not new concepts found in the article set anymore (Webster and Watson 2002). Webster and Watson (2002) also state that effective literature reviews are usually based on a *concept-centric* approach rather than *author-centric* or *chronological* approaches. In this study, both concept-centric and author-centric approaches are exploited.

The previous research of using ML and FS methods in default prediction in P2P lending area is relatively scarce. Therefore, the search of literature is extended to the related fields of study (especially to credit risk assessment and prediction in general) where the research on the use of these models is more developed. Two main streams of literature are included in the review: the first stream is focusing on using ML-based methods and FS in credit scoring and default

prediction in general, excluding the studies related to P2P lending area. The second stream includes the literature concentrating on using the statistical and ML-based methods and FS in the P2P lending credit scoring and default prediction.

5.3 Search process

The initial search was done by using the search of international scientific e-publications in Finna portal. The service is provided by ExLibris and it combines many scientific databases such as Scopus, ScienceDirect, ProQuest and IEEE Xplore in one search engine. This service was chosen to conduct the search because it searches the scientific articles from multiple databases and has also wide variety of different search options. For example, the search can be limited to focus only on the peer-reviewed articles and the keywords can be assigned for instance to full texts, abstracts, or titles.

The keywords were chosen based on the research topic and validated by doing multiple searches with different options. The used keywords associated with P2P lending included “P2P lending” and its synonymous alternatives “social lending” and “peer-to-peer lending”. To include the articles from credit risk area in general in the search, keywords “credit risk”, “credit scoring” and “default” were used. The search of articles in credit risk area was limited to focus on the studies related to ML, FS or prediction by using keywords “machine learning”, “data mining”, “feature selection” and the keyword “predict*” which was formed by using truncation. This keyword helps to include words in the search that contain the exact match of word “predict” but also other forms of the same word beginning with the same character sequence such as “predicting” and “prediction”. Because of the large amount of search results with these keywords, the search was limited to the titles of the articles.

Filtering of relevant articles was conducted using different criteria and limitations. First, the articles that were not peer-reviewed or were not available in full-text versions were removed from the results. Also, to ensure the recency of the research and methods applied in the investigation, the studies that were published before year 2000 were excluded. The literature dealing with the P2P lending in general, different P2P lending platforms or development of the P2P lending markets was also excluded from this review because these topics are discussed earlier in Chapter 2. Furthermore, the literature focusing on the funding success of loan applications was excluded because that is not in the focus of this study. The included studies related to credit risk management in general were limited to the studies that deal with consumer credit risk and therefore the studies focusing on bankruptcy prediction and financial distress prediction of firms were excluded. The search process is illustrated in Figure 16.

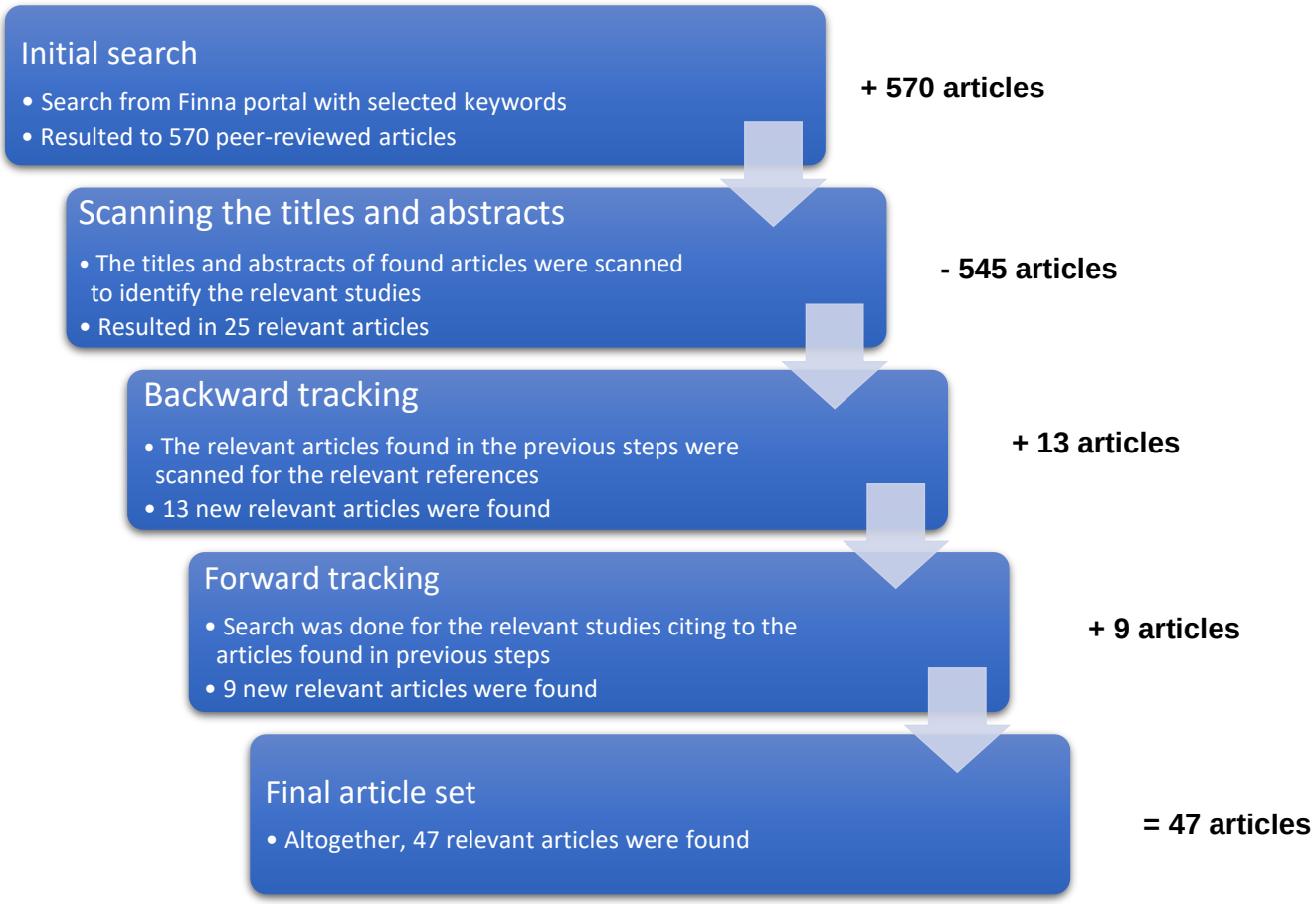


Figure 16. Visualization of the literature search process

After conducting the initial search in Finna service, the titles and abstracts of found articles were scanned through and only the articles relevant to the topic were retained. Next, other relevant articles were searched by using backward and forward tracking and these articles were added to the set of reviewed articles. The process was finished when the topics started to look familiar and new concepts were not found anymore. Through the process, altogether 47 relevant articles were found to be reviewed. It is noteworthy that the reviewed article set is not exhaustive but can be considered as a representative sample of the studies on the topic.

5.4 Statistical and machine learning models in credit risk prediction

As stated by Thomas (2000), wide range of different statistical models have been used in the credit scoring field since the first applications of credit scoring were developed in 1950s. The rapid evolution of the credit card industry and the growing amount of credit clients forced the financial institutions to automate the lending decision processes and accelerated the development of credit scoring models since the 1960s. The statistical credit scoring methods found to outperform the traditional judgemental techniques and helped to reduce the default rates

considerably and this combined with some fundamental regulatory changes encouraged the rise of credit scoring during the next decades (Thomas 2000). The first traditional statistical credit scoring models included for example linear discriminant analysis (LDA) and logistic regression (LR) models (Abdou and Pointon 2011; Thomas 2000).

As the 21st century approached, the growing computational power, cumulative amount of historical consumer credit data and development of more sophisticated statistical techniques made it possible to develop more advanced methods for credit scoring purposes (Abdou and Pointon 2011). The simultaneous growth of artificial intelligence and ML have made it possible to create predictive models that make use of historical data to assess and predict the credit risk of new customers. Until today, the credit scoring models based on the ML algorithms such as decision trees (DT), k-nearest neighbors (KNN), neural networks (NN), support vector machines (SVM), genetic programming (GP) and different kind of hybrid and ensemble methods have been introduced for credit risk prediction (Crook et al. 2007; Louzada et al. 2016).

The used prediction models and FS methods in the reviewed articles are represented in Table 2. To keep the table simple and easy to read, the main objectives of the studies and the used datasets are listed separately in Appendix 1. It is noteworthy that the Table 2 depicts the rough breakdown of different prediction models; different variations of the models are used in the studies, but the models are categorized according to the base algorithms. 4 previous reviews have also been included in the literature review to get comprehensive overview of the topic.

Table 2. Studies related to consumer credit risk prediction in general

	Prediction models							
Author(s) & year	**LR**	**SVM**	**DT**	**NN**	**KNN**	**DA**	**Other**	**Feature selection method(s)**
Wang et al. 2018							GA	F-score, IG, Correlation
Dahiya et al. 2017			X	X				Principal component analysis, Chi-squared test
Butaru et al. 2016	X		X					Prior knowledge / Not reported
Ha and Nguyen 2016					X			SFS, SFFS, MRMR, Random selection, Statistical Dependency, MI
Louzada et al. 2016							Review	Review
Oreski and Oreski 2014				X				Hybrid GA, IG, Gain ratio, Correlation, Gini index
Kruppa et al. 2013	X		X		X			Prior knowledge / Not reported
Oreski et al. 2012				X				GA, SFS, IG, Gini index, Correlation
Abdou and Pointon 2011							Review	Review
Chen and Li 2010		X						DA, DT, F-score method
Khandani et al. 2010			X					Prior knowledge / Not reported
Khashman 2010				X				Prior knowledge / Not reported
Zhou et al. 2010	X	X	X	X	X	X	Adaboost, Probit, BC	Prior knowledge / Not reported
Bellotti and Crook 2009	X	X			X	X		LR, SVM
Chen et al. 2009		X	X				MARS	DT, MARS
Yeh and Lien 2009	X		X	X	X	X	BC	Prior knowledge / Not reported
Yu et al. 2008	X	X		X				Prior knowledge / Not reported
Tsai and Wu 2008	X	X		X				Prior knowledge / Not reported

Crook et al. 2007							Review	Review
Huang et al. 2007		X	X	X			GP	Genetic Algorithm, F-score method
Lee et al. 2006	X	X	X	X				MARS, DA
Liu and Schumann 2005	X		X	X	X			ReliefF, Correlation, Consistency, Wrapper method
Ong et al. 2005	X		X	X			GP, Rough sets	Prior knowledge / Not reported
Somol et al. 2005					X		GC	SFFS, GC, KNN, 3 probabilistic distance measures
Wang et al. 2005	X	X		X			Linear regression	Prior knowledge / Not reported
Galindo and Tamayo 2000			X	X	X		Probit	Prior knowledge / Not reported
Thomas 2000							Review	Review
West 2000	X		X	X	X	X	Kernel density	Prior knowledge / Not reported
Count	**12**	**9**	**13**	**15**	**9**	**4**		**FS methods performed in 12 studies**

New abbreviations used in the table: BC = Bayesian Classifier, IG = Information Gain, GC = Gaussian classifier. For other descriptions, please refer to the list of abbreviations.

Numerous reviewed studies have aimed to propose novel models for credit risk assessment based on different statistical and ML algorithms. For example, Ong et al. (2005) proposed novel credit scoring models based on GP while Lee et al. (2006) based their proposed credit scoring models on DTs and multivariate adaptive regression splines (MARS). Wang et al. (2005) built a new fuzzy SVM for credit risk evaluation whereas Huang et al. (2007) introduced a new SVM-based hybrid credit scoring model. Furthermore, Yu et al. (2008) proposed a new multistage NN ensemble learning approach for credit risk assessment and Zhou et al. (2010) proposed ensemble models for credit scoring based on least squares SVMs. Kruppa et al. (2013), in turn, introduced a new Random forest (RF) based approach for credit default probability estimation. It is worth noting that the hybrid and ensemble methods have gained popularity in the research field in recent years.

The effectiveness of the proposed models has been typically confirmed by comparing the classification results with more traditional models such as DA (Zhou et al. 2010), LR (Kruppa et al. 2013; Lee et al. 2006; Yu 2008), DT (Ong et al. 2005; Lee et al. 2016; Huang et al. 2007) and KNN (Zhou et al. 2010; Kruppa 2013) classifiers. The studies have proven that more sophisticated models frequently outperform the traditional methods in predicting the credit risk, but the traditional models still often offer at least competitive performance (Louzada 2016).

In addition to the studies aiming to propose new models, some studies have also aimed to compare the predictive performance of existing credit scoring models. For example, West (2000) compared the performance of five NN-based credit scoring models and more traditional credit scoring techniques including for example DA, LR, KNN and DT models. He found that the NN-based models typically outperformed the traditional ones and from the traditional models, the LR was found to be the most effective. In contrast, Bellotti and Crook (2000) found that the performance differences between SVM-based classifiers and LR, LDA and KNN classifiers in credit scoring were small and not significant. Moreover, Tsai and Wu (2008) compared the accuracy of NN ensembles to individual NN classifiers in credit scoring and found that the

ensemble classifiers outperformed the individual classifiers only with one of three real-world credit datasets.

Yeh and Lien (2009) compared the performance of KNN, LR, DA, Naïve Bayesian, NN and DT classifiers in credit risk prediction. They found that there was no clear winner among the different models in discriminating the good and bad borrowers, but NN-based models gave the most accurate estimates of the probability of loan default. They also found that KNN classifier performed worse than the other classifiers. In their study, Butaru et al. (2016) found that RF and DT algorithms outperformed the LR model in the prediction of credit card delinquencies.

Also, different FS methods have been under consideration in many previous studies related to credit risk assessment. For example, Somol et al. (2005) compared the accuracy of filter-based and wrapper-based FS methods in credit scoring and found that wrapper-type methods usually outperformed the filter-type methods in classification performance. Furthermore, Liu and Schumann (2005) investigated the effects of different FS methods on model simplicity, model speed and model accuracy in credit scoring. They examined three filter-based FS methods (ReliefF, correlation-based and consistency-based methods) and a wrapper-type method in combination with different classification algorithms. In the study, it was found that consistency-based and wrapper-type FS methods performed better than others. Overall, the investigated FS methods were found to be efficient in selecting the most predictable set of features.

Huang et al. (2007) proposed a hybrid GA-SVM approach for credit scoring in which the GA was used to select the optimal feature set for SVM classifier. The proposed model was tested against the model with F-score-based FS and the model without FS and was found to give equally accurate classification results using markedly less features. Furthermore, Chen et al. (2009) proposed a hybrid SVM-based credit scoring model where the FS was done with MARS and DT-based approaches. The hybrid model with FS based on MARS and DT was found to outperform both the individual MARS and DT classification models and SVM-based classifier without FS.

Moreover, Chen and Li (2010) investigated different combinations of FS methods with SVM-based classifiers in credit scoring. As the FS methods, they used for instance LDA, DT approach and F-score method. It was found that the FS methods were efficient in finding the optimal feature subset in credit scoring. Oreski et al. (2012), in turn, introduced a hybrid GA-NN system for retail credit risk assessment which was tested against several traditional FS methods (Gain ratio, Gini index, correlation and forward selection). The proposed model was found to be the most effective among the tested models. Oreski and Oreski (2014) further modified the GA-NN model by proposing a model that involved preliminary restriction phase

before applying the GA to select the final feature subset. This preliminary phase reduced the initial feature space by using ranking (filter-type FS) algorithms. The novel hybrid GA-NN model was found to outperform the prior GA-NN model.

Ha and Nguyen (2016) used different FS methods (SFS, MRMR, random selection, statistical dependence and MI) with KNN classifier in credit scoring and found that the FS increased the classification accuracy and helped to reduce the complexity of the models. Dahiya et al. (2017), in turn, proposed a hybrid bagging algorithm with FS and tested its performance using both quantitative and qualitative datasets. Principal component analysis was used as FS method in case of quantitative data and the Chi-Square FS in case of qualitative data. The proposed models were found to be efficient in credit scoring. In addition, Wang et al. (2018) proposed a two-phase hybrid FS method that was based on filter approach and multiple population GA. The proposed model was tested against the pure GA FS method and filter-based FS methods with SVM classification algorithm. It was found that the proposed model provided the highest classification accuracy among the tested models.

Different methods have been used to evaluate the predictive performance of credit scoring models in the previous studies. The predictive performance is also typically used to compare the efficiency of different FS methods, but it is remarkable that systematic performance comparisons of different FS methods in credit risk area have been rare. The confusion matrices and measures calculated based on confusion matrix, typically classification accuracy or the classification error (for example, Liu and Schumann 2005; Tsai and Wu 2008; Khashman 2010; Wang et al. 2018), type 1 and type 2 errors (false positive and false negative rates) (Wang et al. 2005; Yu et al. 2008) or type 1 and type 2 accuracies (true positive and true negative rates) (Chen et al. 2009; Dahiya et al. 2017; Lee at al. 2006) have been used in most studies to evaluate the predictive performance. In addition, the AUC measure has been commonly used (for example, Bellotti and Crook 2009; Chen and Li 2010; Khandani et al. 2010). The statistical significance of differences between the different models have been proven typically based on t-test (Oreski et al. 2012; Tsai and Wu 2008; Wang et al. 2005; Zhou 2010) or the non-parametric Wilcoxon signed rank test (Chen and Li 2010; Wang et al. 2005; Wang et al. 2018).

In addition to the classification performance comparison, the performance of FS methods has been investigated for example by comparing the number of chosen features (model complexity) between FS models (Chen and Li 2010; Wang et al. 2018). The datasets used in the studies (see Appendix 1) have mostly been consumer credit and credit card datasets provided by banks. Clearly the most popular datasets have been Australian and German credit datasets which are available in UCI Machine Learning Repository (for example Chen and Li 2010; Huang et al. 2007).

5.5 Credit risk assessment and prediction in P2P lending

The increasing popularity of different P2P lending platforms in the recent years has raised interest in the subject as the research topic. Especially the credit risk prediction and identifying the factors affecting the risk of default have been under investigation. The history of the P2P lending risk assessment and prediction is relatively short. The first P2P lending platform started operating in 2005 and only few studies were published on the 2000s. However, the number of studies has grown rapidly during the 2010s.

5.5.1 Determinants of default in P2P lending

Several empirical studies conducted in P2P lending context have focused on investigating the determinants of P2P loan default. These studies have mostly used traditional statistical techniques to identify the most important variables that affect the loan default probability. The summary of the reviewed studies that have aimed to examine the P2P lending default determinants is represented in Table 3.

Table 3. Studies exploring the determinants of P2P lending credit risk

Author(s) & year	Objective(s)	Data	Statistical model(s)
Polena and Regner 2018	Exploring the determinants of borrower's default in P2P lending	P2P loan data from *Lending Club*	Logistic regression
Lin et al. 2017	Finding the factors determining P2P loan default risk	P2P loan data from *Yooli*	Logistic regression
Dorfleitner et al. 2016	Exploring the effect of soft information on default and funding probability in P2P lending	P2P loan data from *Auxmoney* and *Smava*	Probit regression
Emekter et al. 2015	Credit risk evaluation and exploring P2P loan characteristics	P2P loan data from *Lending Club*	Logistic regression and Cox Proportional Hazard regression
Serrano-Cinca et al. 2015	Finding the factors explaining P2P loan default	P2P loan data from *Lending Club*	Logistic regression
Carmichael 2014	Modeling default of P2P loans and finding the significant factors of default	P2P loan data from *Lending Club*	Dynamic logistic regression
Duarte et al. 2012	Exploring the effect of trustworthiness on credit scores and default rates in P2P lending	P2P loan data from *Prosper*	Proportional hazard model and Logistic regression
Iyer et al. 2009	Exploring the lenders' ability to infer borrowers' creditworthiness from available P2P loan data	P2P loan data from *Prosper*	OLS regression and censored regression

In one of the first studies related to the topic, Iyer et al. (2009) explored the possibility that lenders can infer the borrowers' creditworthiness from the data provided by P2P platform. They analyzed the loan data from Prosper using OLS and censored regression and found that lenders can evaluate part of the credit risk by analyzing both available hard and soft information about the borrower. The impact of soft information on default risk was further supported by the study conducted by Duarte et al. (2012) using the data from Prosper. They found that the borrowers that appear trustworthy default their loans less often than others. Dorfleitner et al. (2016) also examined the role of soft information using loan data from two European P2P platforms and found that the soft information derived from description texts is related to the

funding success. However, they state that the soft information hardly predicts the default probability.

LR and its variations have been widely used in examining the determinants of loan default in P2P lending. Emekter et al. (2015) used LR and Cox proportional hazard regression to examine features that explain the P2P lending credit risk. In their study, they used loan data from Lending Club and found that the credit grade determined by the platform, debt-to-income ratio, FICO score (a credit score assigned by third party to represent the creditworthiness of the borrower) and revolving line utilization were the most important predictors of default. Moreover, Carmichael (2014) used dynamic LR to analyze the defaults of P2P loans made through Lending Club. He found that for instance FICO score, income level and loan purpose were significant factors in explaining loan default whereas the income verification or past bankruptcies did not have statistically significant impact on the loan default.

Furthermore, Serrano-Cinca et al. (2015) studied the determinants of default in P2P lending context using LR model and the loan data provided by Lending Club. In their study, they found that the credit grade assigned by the platform was the most significant predictor of loan default. They state that other information such as the borrower's debt level, annual income and credit history improved the classification accuracy of the model as well. In the study conducted by Polena and Regner (2018), the determinants of default in P2P lending were examined considering the risk class of the loan. Using LR and Lending Club data, it was found that the significant determinants vary between different risk classes and only few of the variables are significant across all the risk classes. According to the results of the study, for example debt-to-income ratio and inquiries in the past half year increased the default probability.

Lin et al. (2017) examined the factors affecting credit risk with LR using the loan data from Yooli, a Chinese P2P platform. Based on the results of the study, the borrower characteristics such as gender, age, educational level, marital status and working years have a significant impact on loan default probability. The study results also showed that for example loan amount and the factors related to borrower's indebtedness such as debt to income ratio and delinquency history explained the default risk.

5.5.2 Credit risk prediction and loan performance evaluation in P2P lending

In addition to more traditional statistical methods, many studies in P2P lending area have exploited different ML and predictive models to predict the P2P lending credit risk and loan performance. The studies related to credit risk prediction and loan performance evaluation in P2P lending have mostly aimed to build an efficient prediction model to predict the P2P lending default or to guide loan decisions made by investors. As was the case in credit risk area in

general, the accuracy of proposed models is typically confirmed by comparing their performance to the models that have been commonly used in the field. Some reviewed studies have also focused on comparing the predictive performance of different models. The reviewed studies are listed in Table 4. It is noteworthy that different variations of the listed prediction models are used in the studies, but the models are categorized according to the base algorithms. The FS models are also used in many studies to enhance the prediction accuracy of the models but as it can be seen from Table 4, the FS has frequently been done by using the prior knowledge of the topic or it is not reported in the studies at all.

Table 4. Studies related to risk assessment in P2P lending

			Statistical / machine learning model(s)						
Authors & year	**Objective(s)**	**Data**	**LR**	**SVM**	**DT**	**NN**	**KNN**	**Other**	**Feature selection**
Teply and Polena 2020	Comparing the classification models in P2P lending	P2P loan data from *Lending Club*	**X**	**X**	**X**	**X**	**X**	BC, LDA	Prior knowledge / Not reported
Chen et al. 2019	Assessing the probability of default and significant impact variables	P2P loan data from *PPDai*	**X**						Stepwise AIC, LASSO, Bayesian variable selection
Zhou et al. 2019	Default prediction in P2P lending	P2P loan data from Chinese platform	**X**	**X**	**X**	**X**	**X**	AdaBoost	LMBFR
Ye et al. 2018	Loan evaluation in P2P lending	P2P loan data from *Lending Club*	**X**	**X**	**X**		**X**		Prior knowledge / Not reported
Xia et al. 2017	Creating a portfolio allocation model and predicting default risk in P2P lending	P2P loan data from *Lending Club* and *We.com*	**X**		**X**				XGBoost feature importance score and SFS
Guo et al. 2016	Credit risk assessment in P2P lending	P2P loan data from *Lending Club* and *Prosper*	**X**					Instance-based model	Prior knowledge / Not reported
Zhang et al. 2016	Credit scoring in P2P lending	P2P loan data from *PPDai*	**X**		**X**	**X**			Prior knowledge / Not reported
Serrano-Cinca and Gutierrez-Nieto 2016	Comparing the performance of profit scoring and credit scoring models in P2P lending	P2P loan data from *Lending Club*	**X**		**X**			Linear regression	Prior knowledge / Not reported
Byanjankar et al. 2015	Credit scoring in P2P lending	P2P loan data from *Bondora*	**X**			**X**			Prior knowledge / Not reported
Jin and Zhu 2015	Default risk prediction in P2P lending	P2P loan data from *Lending Club*		**X**	**X**	**X**			LMBFR and correlation
Malekipirbazari and Aksakalli 2015	Credit risk assessment in P2P lending	P2P loan data from *Lending Club*	**X**	**X**	**X**		**X**		Predictive power on response variable
Count			**10**	**5**	**8**	**5**	**4**		**FS performed in 4 studies**

New abbreviations used in the table: BC = Bayesian Classifier, AIC = Akaike's information criteria. For other descriptions, please refer to the list of abbreviations.

For example, Malekipirbazari and Aksakalli (2015) proposed a RF-based method for credit risk assessment in P2P lending. They used the predictive power of the features on response variable as the FS method and found that the RF model outperformed other tested classifiers (SVM, LR and KNN) in identifying good borrowers. Furthermore, Jin and Zhu (2015) compared the predictive accuracies of five different classifiers in P2P lending using the predictor importance score of RF classifier and correlation matrix as FS methods. Two of the tested

classification models were DT-based, two were NN-based and one was SVM-based. The results of the study showed that the SVM-based classifier slightly outperformed the other models.

Byanjankar et al. (2015) proposed a NN-based model for credit scoring in P2P lending and found that the NN model outperformed the LR model in screening defaulted loans. Zhang et al. (2016) built a DT-based credit scoring model where the social media information was used as the part of classification model. In the study, it was found that the DT-based classifier outperformed the LR and NN classifiers in predicting the P2P loan default.

The profit scoring -based models have also been used more and more often in recent years in P2P lending loan evaluation. For example, Serrano-Cinca and Gutierrez-Nieto (2016) introduced a novel profit scoring system based on multinomial linear regression and DT and found that it outperformed the LR credit scoring model in maximizing the expected profitability (internal rate of return) of loan portfolio. Furthermore, Guo et al. (2016) proposed an instance-based model for credit risk assessment in P2P lending and found that the proposed model outperformed the LR-based model both in default probability and return rate prediction.

Moreover, Xia et al. (2017) built a cost-sensitive boosted tree model for P2P loan evaluation to enhance the efficiency in discriminating potential default borrowers. They used FS method combining feature importance scores of the boosted tree and SFS to select the optimal feature subset. It was found that the extreme gradient boosting decision tree model outperformed LR-based and RF-based models both in estimating the profitability of P2P lending investments and discriminating the bad borrowers from good borrowers in most cases. Ye et al. (2013) proposed a P2P loan evaluation model which was based on RF classifier optimized by GA with profit score. The proposed model was proven to lead to higher profit for the investor than the conventional models (RF, SVM, DT, KNN and LR).

Even if the profit scoring has lately got attention on loan evaluation research field, the loan default prediction is still a common research topic in P2P lending area and new models are still developed for the default prediction purposes. For example, Zhou et al. (2019) proposed a new DT-based ensemble default prediction model for P2P lending which integrated three DT-based classifiers (gradient boosting decision trees, extreme gradient boosting decision trees and light gradient boosting machine). In the study, the FS was conducted using learning-model based feature ranking. The proposed ensemble model was found to outperform the individual DT-based models as well as NN, LR, RF, SVM, KNN and Adaboost algorithms in default prediction.

In addition, Chen et al. (2019) used logistic quantile regression to predict the default risk and in P2P lending. As the FS methods, they used Stepwise AIC method, LASSO and Bayesian

variable selection. It was found that the Bayesian method performed better than two other methods in selecting the most discriminating features. Teply and Polena (2020), in turn, constructed a systematic ranking of 10 different classification algorithms in P2P lending. They found that LR, ANN and LDA were the most accurate algorithms followed by linear SVM and RF models. KNN and DT were found to be the two worst classifiers in predicting the loan default.

Also, the significant determinants of P2P lending default have been under investigation in many studies that have aimed to predict the defaults of the loans. Both statistical measures and learning-model based feature importance estimates have been used in investigating the relative importance of used features. For example, Malekipirbazari and Aksakalli (2015) used information gain and correlation as measures and found using Lending Club data that the loan grade assigned by the platform, income to payment ratio, annual income, FICO score and debt to income ratio had the strongest impact on loan default. Also, Jin and Zhu (2015) used the Lending Club data and found using the relative importance of predictors from different classification models as the measure that the loan term, annual income, loan amount, debt to income ratio, credit grade and revolving line utilization had the most significant impact on loan default. The findings of the study conducted by Ye (2018) supported mostly the previous results obtained using the Lending Club data.

Moreover, Xia et al. (2017), measured the relative importance of features using the importance scores of extreme gradient boosting DT and found that also the nominal interest, home ownership and employment length were significant determinants of default in the Lending Club data. In the data provided by Chinese P2P platform We.com, the most significant determinants were loan term, nominal interest rate, number of passed verifications, educational attainment, and employment length. Using the data from Chinese P2P platform PPDai, Chen et al. (2017) found that the loan period, interest rate, loan type and regulationary changes affected the default probability of the P2P loans the most. Byanjankar et al. (2015) used relative importance of the NN-based classification models as measure and found using the data from Bondora that the binary variable indicating whether the underwriters had restructured the initial loan application, the applied loan amount, total income and country of the borrower had the highest impact on the model outcome.

The evaluation of the classification and FS models in P2P lending context has been conducted mostly with similar methods than in the case of credit risk assessment and prediction in general. However, it is noteworthy that the FS methods are systematically compared to each other only in one reviewed study. Overall classification accuracy (Malekipirbazari and Aksakalli 2015; Ye et. al 2018; Zhang et al. 2016), other confusion matrix-based measures (Byanjankar

et al. 2015; Ye et al. 2018; Zhou et al. 2019) and AUC measure (Xia et al. 2017; Zhou et al. 2019; Teply and Polena) are the most used measures of classification performance. In case of profit scoring, the profitability-based evaluation methods have been exploited. The datasets used in the studies include loan datasets provided by different P2P lending platforms. Clearly the most popular dataset is the loan dataset provided by US platform Lending Club (for example, Emekter et al. 2015; Serrano-Cinca and Gutierrez-Nieto 2016; Teply and Polena 2020). Most of the datasets have been provided by US and Chinese P2P platforms.

5.6 Summary of the literature review

This literature review was conducted to get an overall look on the previous research related to credit risk assessment and prediction in general and more specifically in P2P lending area. The first research question of this thesis and the corresponding sub-questions can be answered based on the conducted literature review. They were formed as follows:

1. What is the current state of credit risk assessment and prediction in the scientific literature?
    a. What statistical and machine learning models have been the most popular in credit risk assessment and prediction in previous studies?
    b. How has the feature selection been performed in previous studies and how have the methods been evaluated?
    c. What variables have been found to explain the credit risk in P2P lending in the previous research?

The answer for the first sub-question (1a) is that from the ML models, the models based on NN, LR, DT, SVM and KNN classifiers are used most frequently in credit risk assessment and prediction in general. In the P2P lending area, the same ML models have been widely used. From the statistical models, the LR and its variations are used most often. In the recent years, more and more sophisticated ML classification models have been introduced and commonly used in credit assessment and prediction, and the ensemble and hybrid methods have become more popular. In P2P lending area, especially the RF model has been popularly used from the ensemble classification methods. Even if more sophisticated models have been developed in recent years, the conventional models (especially LR) are still commonly used in credit risk assessment and prediction.

To answer the second sub-question (1b) it can be stated that the FS in credit risk assessment and prediction has been done using wide variety of different methods. Unlike in case of used ML methods, there are no established state-of-art FS methods in the research field. Filter-

based, wrapper-based as well as embedded and hybrid methods have been exploited in previous studies. Different kind of classifier-specific predictor importance scores and sequential feature selection methods have been used in several studies. From the filter-type methods, the dependency measure-based methods such as correlation and information gain have been used most often.

Even though different kind of automated FS methods are widely used in the previous reviewed studies, it is also worth noting that the FS has often been done using the previous knowledge and expertise of the topic or the FS methods are not reported in the studies at all. The used FS methods have been evaluated mostly based on the final classification performance but also the model complexity (number of features in the proposed feature set) and the modeling speed (the training time of the model) have been used as criteria for the model comparison purposes in some studies.

As the answer for the third sub-question (1c) it can be said that the risk ratings provided by P2P lending platforms have been found to strongly affect the borrower's default risk. Also, borrower's demographics (for example gender, age, education, and marital status) and credit history (for example the previous inquiries and revolving line utilization) have been proven to explain the loan default according to the previous literature. In addition, the income level of the borrower and loan characteristics such as the loan amount and the purpose of the loan have been found to affect the default probability in P2P lending. Also, the soft information such as the information obtained from social media have appeared to determine the default probability according to the previous research findings.

6 EMPIRICAL ANALYSIS AND RESULTS

In this chapter, the empirical analysis of this study is reported. The process of conducting the empirical part is illustrated in Figure 17.

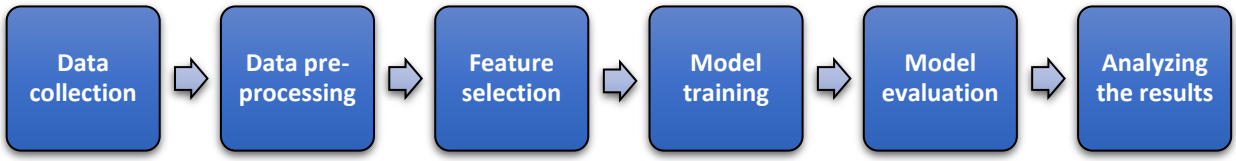


Figure 17. Process of the empirical part of the study

In the following subchapters, the data used in the analysis is first introduced and the preprocessing phase is described. Then, the FS and model training steps are explained. That is followed by the explanation of model evaluation phase. Finally, the obtained results are analyzed in detail and the second and third research questions of the thesis are answered.

6.1 Data collection and pre-processing

The dataset used in this study is provided by Bondora, a well-known Estonian P2P lending platform which was introduced in more detail in Chapter 2.2. The platform provides historical loan data of the loans made through it and offers open access to this daily updated dataset through their websites (Bondora 2019). Another P2P loan dataset – the loan dataset from Lending Club which is nowadays the world's biggest P2P platform – was considered as well but the data restriction policy maintained by the platform forced to exclude this dataset from this study. The Lending Club data is only accessible for the users of the platform, and the registration is possible only for US residents.

The Bondora loan dataset consists of data of all the loans applied through the platform that are not covered by the data protection laws (Bondora 2019). The dataset was downloaded from Bondora websites on 23rd January 2020, and it covered initially information of over 130 000 loans from years between 2009 and 2020. Initially, the dataset included 112 features.

Because the dataset was initially large and incomplete, a lot of preprocessing was needed before it could be used for the modeling. The purpose of preprocessing phase was to clean and preprocess the data so that it could be used for analysis with selected classification and FS methods. To make the FS process reliable and to ensure that the most significant features were really selected by the FS methods themselves without the major effect of user's prejudices on the selection, the original feature set was kept as large as possible. However, many of the features had to be removed because they had a lot of missing values or could have caused the data leakage in the model. Because of the large size of the dataset, the

preprocessing phase was conducted with RStudio software and selected additional R packages. For example, the *dplyr* and *purrr* packages were used which are included in the collection of R packages called t*idyverse*.

6.1.1 Handling missing values and initial variable removal

There were initially large number of missing values associated with some features that have been found significant for the default prediction in the previous research. These variables did not have any values either before certain point in time (many variables started to have values in November 2012) or after certain point (especially after June 2017) in the dataset. For example, the values of the variables indicating the debt-to-income ratio, existing liabilities, and the type of the home ownership of the borrower were missing in the beginning of the dataset but had the values afterwards.

The variables which did not have any values after certain time point in the dataset included for example educational level, employment status and marital status. Also, the variables associated with the use of the loan and the features related to different income streams of the borrower had no values after a certain point in time but had values before that.

The limited availability of the values of these variables forced the timespan of the dataset to narrow to the years from 2012 to 2017. However, removing the very first observations from the data can even make the analysis more reliable because the business took some time to stabilize after the start of its operation (Bondora started operating in 2012). The very first observations might have distorted the results because for example, the share of the defaulted loans was considerably smaller in the first years of operating compared to the rest of the sample. Also, the latest years would have been excluded from the dataset anyway because the latest loans have not had much time to default yet.

Many features were also excluded from the initial dataset in the pre-processing phase since they do not have much to do with creditworthiness. Variables such as “Loan ID”, “Loan number” and “UserName” were removed from the data since they are assigned to each loan (or borrower) mainly for data storage and identification purposes and are therefore considered meaningless for default prediction. The date variables that were considered unnecessary (for example the timestamp of the loan application appearing in the primary market) were also excluded from the dataset.

Also, the variables that could have potentially increased the problem of data leakage were removed from the dataset. The data leakage can be introduced to the model if features that would not be available in practice when the model is used for the prediction or the features

that actually have the same information that the predictive model is trying to predict are used to train the model. That can lead to biased estimates and over-optimistic predictive models (Kaufman et al. 2012).

For this reason, all the features that are not available at the time when the investor is making the investment decision (whether to invest in a certain loan or not), were excluded from the final dataset. These variables included, for example, information related to payments during the loan duration and the variables determining the overdue payments. The old versions of Bondora's credit ratings, the variables representing the estimated values for expected loss and return of the loan and the variable indicating the probability of default were also removed from the dataset to prevent the data leakage problem. It is noteworthy that the number of variables with the possible data leakage problem in the initial dataset was considerable. Excluding these variables from the analysis led to the removal of 52 variables from the dataset.

In addition, the categorical variables presented in character format that had a very large number of different levels (such as the variables indicating the county and city of the borrower) were excluded from the dataset. In the dataset, there are borrowers from many different countries so the cardinality of these variables would have been very high. Among the problematic categorical variables was also the variable indicating the employment position of the borrower. In this variable, there would have been a large number of different classes and the descriptions were also given in multiple languages (in English and in Estonian) which would have made the variable complicated to analyze.

Four variables representing credit scores provided by different third-party credit rating institutions had initially a lot of missing values. Typically, each borrower was assigned only one of these scores and other scores were missing. However, the number of observations with all the four scores missing was notably lower. The number of missing values of each credit score are shown in Appendix 2. To make it possible to use this information in the analysis, different scores were rescaled on the same scale (from 1 to 6) and combined into one variable "Credit score". If only one of the four credit scores was available, that was used alone. Otherwise, the average of available credit scores was calculated. Because of the rescaling and average calculations, the credit score is later handled as a continuous variable.

Finally, to make the dataset complete and to make it possible to use the dataset with ML models, the incomplete observations were removed from the dataset using the complete case analysis. Other ways of handling missing values were considered as well but the direct removal was chosen because the amount of incomplete observations was relatively small, and the

missing values were found to be random by nature. The final dataset still includes about 30,000 observations which can be considered as a sufficient sample for the purposes of this thesis.

6.1.2 Encoding of categorical features

Initially, there was no variable in the dataset directly indicating the default of the loan, so the target variable was encoded from the variable “Default date”. This variable initially determined the date when the loan default occurred, and the collection process was started. Therefore, the presence of reported default date indicated that the default had occurred and if no default date was reported, no default had happened in case of a corresponding loan. The binary target variable was named “Actual default” and it takes the value 1 if the default date is reported (the default has happened) and 0 otherwise (if no default has occurred).

The data initially included many categorical variables in character format which had to be encoded into numerical form before modeling. Both label encoding and One Hot Encoding were used as encoding techniques in case of different FS methods, depending on the characteristics of FS methods. In case of different classification models, Matlab’s in-built procedures were used to handle the categorical variables.

6.1.3 Outlier removal and handling the high cardinality of categorical variables

Because ML algorithms are frequently prone to outliers, the dataset was further examined for strong outliers. Especially in the case of variables measuring the total income of the borrower and different sources of borrower income, there were some observations with exceptionally high values. Initially, there were four instances in the dataset with total monthly income of over 60 000€ which leads to over 700 000€ yearly income. Three of four of these borrowers reported over 100 000€ monthly income. These four observations were removed from the dataset because they had extremely high values also in case of other variables. In case of income from leave pay there was one extremely high value. One borrower reported the monthly income from leave pay to be over 20 000€. This observation was also excluded from the final dataset.

Some categorical variables had initially very small number of observations in some of the categories (some of the class frequencies were very low). This could have been problematic because for example estimating the LR coefficients for variables with very few observations can make the classification results less reliable. The variables with high cardinality included loan duration, home ownership type and monthly payment day. In the case of loan duration, the durations of 1, 2, 4 and 5 months had less than 10 observations each. Furthermore, home ownership type class “0 = homeless” had only one observation and the monthly payment day “28” had less than 10 observations. All of these observations were removed from the final

dataset (the categories with very low class frequencies were removed from the categorical variables).

6.1.4 Data split

The data was split into training and test sets using simple holdout validation approach. 70% of the data was used as the training set and the remaining 30% was used as the independent test set for evaluating the out-of-sample performance. The simple holdout method was chosen to split the data initially because the cross-validation (CV) approach would have forced to use the *nested CV* (also called *double CV*) procedure in the model selection phase to ensure the independence of the test set (Suppers et al. 2018). This would have increased the computational time considerably. Also, the dataset can be considered large enough to hold out the independent test set in the first place. The initial split was done with stratified sampling technique which ensures that the classes of target variable are adequately represented in training and test sets (Kohavi 1995). In this case it means that the share of defaulted and non-defaulted instances in training and test sets are approximately same as in the whole dataset.

Using simple holdout method makes the final classification results dependent on one random split into training and test sets. This can lead to biased results if the training and test sets are not representative samples of the initial data. This bias is often reduced by splitting the data randomly into training and test sets multiple times, training and testing the model with different training and test sets and averaging the classification results. However, splitting the data multiple times with holdout method would lead to over-optimistic classification results because the instances used in model selection and final classification performance evaluation phases would be partially the same (Kohavi 1995). Therefore, the single holdout split was considered as a sufficient approach. The representativeness of the training and test samples is illustrated later in Chapter 6.2.2 in connection with the descriptive statistics.

5-fold CV was used to validate the FS and hyperparameter optimization phases. This helps to make the FS and hyperparameter optimization results less biased and makes it possible to exploit the whole training data in model selection phase (Arlot and Celisse 2010). The basic idea of CV approach used in this study was described earlier in Chapter 3.5.

6.1.5 Data standardization

The dataset had initially numerical variables that were measured on different scales. For example, the values of the total monthly income (after removal of the most significant outliers) ranges from 200€ to 17000€ whereas the number of previous loans ranges from 0 to 23. Very different scales of the variables can markedly distort the results of the classification and affect

the results of FS phase especially in the cases where the distance measures are used as the part of the algorithms. That is why all the variables except from the categorical variables were normalized with min-max normalization technique. The formula of the normalization can be represented as follows (Aksoy and Haralick 2001; Patro and Sahu 2015):

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

In the formula, z_i denotes the normalized value and x_i is the corresponding instance of variable x. Therefore, the minimum value of the variable was assigned to take value 0, maximum value was transformed to 1 and all the other values were transformed to be between 0 and 1.

6.2 Descriptive statistics

The final dataset after pre-processing contains 46 variables (including the target) and 29 375 observations. The class frequencies of the target variable "Actual default" are represented in Table 5. It is noteworthy that the default rate of the loans made through Bondora under the period examined has been relatively high: slightly above 56% of the loans have been defaulted. This can be due to the adverse selection and moral hazard problems that lead to the situation where borrowers with high default risk are more willing to borrow money through P2P lending platforms. The credit approval policies maintained by the platforms also strongly affect the default rates in P2P lending platforms (Polena and Regner 2018). Because the class frequencies of the target variable were somewhat equal, there was no need for balancing the dataset.

Table 5. The class frequencies of the target variable

Class	0 = Non-default	1 = Default	Overall
Frequency	12893	16482	29375
Relative frequency	43.9%	56.1%	100.0%

The final dataset includes 45 features. The summary of the features used in the study is represented in Appendix 3. The table gives an overall look of the used features with brief descriptions. In the table, similar breakdown of the features into different groups has been used as in the study conducted by Serrano-Cinca et al. (2015). As it can be seen, the used features are indicating for instance the credit rating of the borrower and the overall characteristics of both borrower and the applied loan. Also, the features indicating the income information, the credit history and the indebtedness of the borrower are under investigation. It is noteworthy that 20 of 45 features in the dataset are categorical (the share of categorical variables in the dataset is relatively high). This strongly affects the selection of used FS and classification methods.

The descriptive statistics of the numerical variables are represented in Appendix 4. As it can be seen, the average of the applied loan amount has been about 3085€ under the period examined. The average total monthly income has been about 1385€ and it has come mainly from the work: on the average, about 87% of the total income has come from principal employer. The age of the borrower ranges from 19 to 72 with the average of 38. Borrowers have on average about 856€ of total liabilities and the free cash available after monthly liabilities is on average about 470€ per month. The debt to income ratio has been on average about 31%.

It is noteworthy that the maximum value of the maximum accepted interest rate is very high (262.63%) and there are over 700 loans where the maximum accepted interest rate is determined to be over 100%. This can indicate that people are negligent when filling out the loan application or that they are desperately seeking for a loan with any interest rate. The mean of the maximum accepted interest rate is 35.84% which can be considered relatively high, reflecting the higher risk of the P2P lending compared to traditional lending. The median of maximum accepted interest rate is over 5% lower than the mean (30.13%) which can be considered relatively big difference. This is at least partially due to the fact that there are a lot of very big values in this feature as mentioned earlier.

It is also worth noting that the distributions of most continuous variables are skewed or leptokurtic. The skewness and kurtosis values for each continuous variable are represented in Appendix 4 and the distributions of continuous variables are visualized in Appendix 5. The results indicate that especially the variables measuring the income information and the information of the credit history of the borrower are strongly skewed to the right. Excess skewness and kurtosis can affect the results of the analysis conducted with ML algorithms because many models assume that the variables are normally distributed. However, the logarithmic transformation conducted for skewed variables did not affect the results notably, so the final analysis is done without logarithmic transformations.

The class frequencies of the categorical variables are represented in Appendix 6 and the distributions are further visualized with bar charts in the Appendix 7. To keep the frequency table simpler, the variables indicating dates are excluded from the table. However, the distributions can be examined from the bar charts also in the case of dates. As mentioned earlier, the high cardinality of categorical variables was handled by removing the observations with very rare class labels (with less than 10 observations). After these corrections, there are still several variables with relatively low class frequencies. These variables include verification type, language code and employment status. However, all the remaining categories have at least 70 observations and the problem of imbalanced categorical features was not considered to be critical for the results of the FS and classification.

The class frequencies indicate that over 65% of the loan applicants are new credit customers. About 55% of the borrowers are Estonian, the rest are either Spanish or Finnish. The gender distribution is relatively even but men have borrowed slightly more: about 53% of the borrowers are men, 40% are women and the rest are unknown cases. Home improvement is the most common loan purpose (excluding the “other” class) with slightly less than 27% share of the whole data. The most common educational level is the secondary education with about 38% of observations and most of the borrowers are fully employed (about 82% of the data). The longest loan duration (60 months = 5 years) is the most popular covering almost half of all observations. The most common risk class (determined by Bondora) is “HR” (the highest risk class) with about 24.19% of the observations, further indicating the high-risk nature of loans.

6.2.1 Statistical dependence analysis

Because the aim of the classification is to classify the observations into different classes of the target variable, it is useful to examine statistically the relationships between features and the target classes. The following analysis is conducted with the whole dataset to tentatively examine the statistical relationships of variables. Because the target variable is binary, the relationships between continuous predictors and the target variable are measured by calculating point-biserial correlation coefficients. The point-biserial correlation is an alternative for Pearson’s correlation coefficient and is commonly used correlation measure when one of the investigated variables is dichotomous (Serrano-Cinca et al. 2015).

The results are represented in Appendix 8 and they indicate that the credit score assigned by third party credit rating agency and the maximum interest rate accepted by the borrower have the strongest positive linear relationships with the loan default in case of continuous variables. Also, the applied loan amount and the income level of the borrower are in a relatively strong positive relationship with loan default. It is logical that the bigger loan amount seems to lead to higher probability of default but the total income’s positive relationship with loan default probability can be considered surprising. Contrarily, the previous repayments before loan has the strongest negative relationship with the default class. Also, the number and amount of previous loans seem to decrease the default risk based on the point-biserial correlation.

Because the calculation of correlation coefficients is not meaningful in the case where all the investigated variables are categorical, the relationships between categorical variables and the target variable are measured by conducting the Chi-Square tests of independence. The results are represented in Appendix 9 and they indicate that the credit rating assigned by Bondora has the strongest relationship with the target variable. There seems also to be a strong association between the country of the borrower and the target variable. Furthermore, the monthly

payment day and the loan duration seem to be relatively strongly associated with the target variable.

6.2.2 Training and test sets

To ensure the representativeness of training and test sets, it is important that the class frequencies of the target variable in the samples are similar than in the whole data. The class frequencies of the target variable in training and test data are represented in Table 6 and as it can be seen, the ratios of defaulted and non-defaulted loans (expressed to one decimal place) are the same in both sets.

Table 6. Class frequencies of target variable in training and test data

	Training data			Test data		
Class	**0 = Non-default**	**1 = Default**	**Overall**	**0 = Non-default**	**1 = Default**	**Overall**
Frequency	9025	11538	20563	3868	4944	8812
Relative frequency	43.9%	56.1%	100.0%	43.9%	56.1%	100.0%

To further investigate the representativeness of the training and test data, a few features with a strong relationship with the target variable were selected and basic descriptive statistics for these variables were calculated in both datasets. The statistical dependence analysis conducted in previous chapter was used to select the important features.

In case of continuous variables, 2 variables with the highest point-biserial correlation coefficient values (credit score and maximum interest rate) were chosen for the analysis. The descriptive statistics for these variables (after standardization) are represented in Table 7. The results show that the means and standard deviations are very similar across training, test, and whole dataset. This indicates that the training and test sets are distributed similarly than the initial dataset.

Table 7. Descriptive statistics of continuous variables in training and test data

	Credit score				Interest rate			
Variable / data	**Min.**	**Max.**	**Mean**	**STD**	**Min.**	**Max.**	**Mean**	**STD**
Whole data	0.000	1.000	0.220	0.244	0.000	1.000	0.114	0.103
Training data	0.000	1.000	0.222	0.245	0.000	1.000	0.115	0.104
Test data	0.000	1.000	0.217	0.243	0.000	1.000	0.113	0.100

According to Chi-Square tests, two most important discrete features (credit rating and country) were chosen to be analyzed. The shares of non-defaulted and defaulted loans across different classes of the features are represented in Appendix 10. As it can be seen from the table, the shares are very similar across different classes of investigated variables, and this further indicates that the training and test samples are representative samples of the whole data.

6.3 Justification of used methods

In this thesis, 4 different feature selection methods are tested in the P2P lending default prediction. The chosen FS methods are:

1. Maximum-relevance-minimum-redundancy (MRMR) approach
2. Chi-Square feature selection
3. Sequential forward selection (SFS)
4. Learning-model based feature ranking (LMBFR)

Matlab software package (version 2020a) was used as a tool for conducting FS in this study. The used models were selected based on the suitability for the used data and their availability in the selected software package. Also, the popularity of the models in previous research on the credit scoring field and in ML field in general was considered in selection.

MRMR FS approach was chosen because it is commonly used in ML field and has been proven efficient with different datasets. The MRMR FS method has been widely applied especially in gene expression (Ding and Peng 2003), but it has been used also in credit scoring area (Ha and Nguyen 2016). The mutual information (MI) as the evaluation criterion is also well-suited for cases where the dataset includes many categorical variables and therefore Matlab's MI-based MRMR algorithm is a suitable choice for feature selection in case of Bondora dataset.

FS method based on Chi-Square tests was chosen to be used because it is applicable for the FS with mixed types of variables (both categorical and continuous variables). It also offers an intuitive and computationally efficient way to perform FS. The Chi-Square FS has also been used earlier in credit risk research area (Dahiya et al. 2017).

SFS has been used in wide range of research fields, and it has been exploited also in the previous studies in credit scoring area (Oreski et al. 2012; Somol 2005). In addition, the SFS method has also been used earlier in P2P lending context (Xia et al. 2017). The SFS method has been proven efficient in terms of the classification accuracy and as a wrapper method the algorithm also exploits the interaction with the classifiers. It is also interesting to compare the performance of FS methods based on different search heuristics, and due to that, one pure wrapper-based method is also included in the analysis even if the wrapper methods are computationally more expensive than the filter methods (Kohavi and John 1997).

LMBFR method was selected because it has been used in many previous studies in P2P lending area (Jin and Zhu 2015; Xia et al. 2017; Zhou et al. 2019), and it offers a simple but effective way to select the important features for classification model. Because the classification model itself is used to estimate the feature importance scores, the feature set can be expected to be

efficient with the corresponding classifier. It is also interesting to examine the performance of an embedded FS method in the context of the research topic. The LMBFR method was used only with DT and RF classifiers because the estimation of relative importance of features is less straightforward for other used classifiers (Xia et al. 2017).

Different FS methods are tested in this thesis in combination with 4 classification models. The chosen classification models are:

1. Naïve Bayes (NB)
2. Logistic regression (LR)
3. Decision tree (DT)
4. Random forest (RF)

All the used classification models have been used in many fields of research and can be seen to belong to the state-of-art ML algorithms. The used models are also available in Matlab software package and are relatively computationally light.

NB classifier was chosen to be used in this study because despite its simplicity, it has proven to be accurate in various classification tasks. It is also efficient in terms of computation time and offers a good benchmark for more sophisticated classification methods (Provost and Fawcett 2013, p.241). The NB classifiers have also been tested earlier in P2P lending research area (Teply and Polena 2020).

Based on the literature review of this study, LR is the most used ML model in P2P lending context. It has also frequently offered at least comparable performance compared to more sophisticated models. For example, Teply and Polena (2020) found that the LR model performed best among the 10 tested classification models in P2P lending default prediction. The LR model is also relatively easy to interpret and is considered as one of the state-of-art techniques in credit scoring area. Furthermore, the LR classifier is well suited for binary classification. For these reasons, the LR model is chosen to be used also in this study.

DT model was chosen because its interpretability and popularity in ML area. The DT model and its variants have also been frequently used in credit risk prediction (Butaru et al. 2016; Chen 2010). In addition, the DT classifiers have been applied in P2P lending research area (Teply and Polena 2020; Zhang et al. 2016). The DT is also well applicable to the classification problems where one or more of variables are categorical and is relatively fast to implement.

Because the popularity of different ensemble models has increased notably in the credit risk prediction in the recent years, one ensemble model (RF) was also chosen to be used in this

study. The RF model has been frequently used also in the P2P lending context, and based on the literature review of this thesis, the RF classifier is among the most frequently used ensemble methods in P2P lending default prediction (Malekipirbazari and Aksakalli 2015; Zhou et al. 2019). The RF model has also provided relatively good performance in many previous studies, and for example in the study conducted by Malekipirbazari and Aksakalli (2015), the RF classifier performed best from the tested classification models.

6.4 Feature selection

The FS methods used in this study include different types of FS methods. The MRMR and Chi-Square FS methods are basically filters, SFS method is a wrapper and the LMBFR method is basically an embedded method. The final feature subset selection procedures of different types of FS methods differ from each other and the selection of the final feature subset can often be done in many ways. Therefore, the principles used in this study are discussed briefly before representing the final FS results of each method.

Because the filter methods aim to rank the features according their importance but do not propose an exact optimal subset of features, the final decision of the chosen features is left to the user. One possible way to do the final decision is to set a fixed number of features (x) and include x features with the best rankings to the final feature subset. Alternatively, a minimum threshold for the value of used evaluation criterion can be defined. Then, only the features with the evaluation higher than the threshold can be used in final classification (Chandrashekar and Sahin 2014). However, determining the exact number of features or the minimum threshold for the evaluation criterion beforehand can be difficult, and the approaches do not guarantee the maximum classification performance (Saeys et al. 2007).

Therefore, the final decision of the used features in case of filter-type FS methods in this study is done based on the CV performance with each classifier. The features are added sequentially to the classification models according to the order proposed by used FS algorithm and the CV error is calculated with each feature combination. Then, the number of features (feature subset) that maximizes the CV performance (or provides a competitive accuracy with fewer features) is selected to the final classification. Similar approach for final decision of features has been used earlier for example by Liu and Schumann (2005).

In case of wrapper-type SFS algorithm used in this study, there are also a few possible ways to make the final decision of the used features. In one possible approach, pre-determined stopping criterion is used to select the final feature subset. Frequently, the search is forced to stop when no improvement in the objective function (usually in the value of classification

accuracy or classification error) can be obtained by adding any of the remaining features in the feature subset. However, this approach does not guarantee the selection of optimal feature subset because the algorithm can get stuck in local optimum of the objective function. (Liu and Motoda 2007, p.24). Another way to decide the final number of features is to set a fixed number of features that should be included in the final subset. Nevertheless, determining the fixed number beforehand is again problematic (Liu and Schumann 2005).

To solve these problems, the SFS algorithm can be forced to continue the search until all the features are selected into the feature set. If adding any of the remaining features does not improve the value of the objective function, the feature that leads to the least reduction in objective function value is selected. The value of the objective function can be used directly as the evaluation criterion: the number of features that minimizes the objective function value is considered as optimal (Liu and Schumann 2005). In this study, this approach is used.

LMBFR method was applied only with DT and RF classifiers. The procedure was started by estimating the relative feature importance scores using the corresponding classification model. Then, the importance scores were used to rank the features according to their importance in classification task. After that, the features were added sequentially to the model according the ranking and CV error was used to select the optimal feature subset in the same way as in case of filter-type FS methods.

It is noteworthy that the final selection of the number of used features with filter-type (MRMR and Chi-Square) FS methods and embedded-type LMBFR method was done based on the CV error and thus the classifiers were incorporated in the final selection procedure. This makes these FS methods technically to be wrappers. However, the effects of using different FS methods on the classification performance can still be compared to each other but the differences in model performance actually reflect the impact of using different logics to construct a feature ranking in the first place. This should be taken into account when analyzing the final results.

Matlab's default values for the model hyperparameters (see Chapter 6.5 for closer details) are used for the NB, LR and RF classifiers in the FS phase. However, in case of DT, the overfitting problem was considered significant and therefore *minimum leaf size* hyperparameter (described in more detail in Chapter 6.5.3) was optimized using Bayesian optimization for each model (Mantovani et al. 2018).

6.4.1 Filter-type feature selection

Two filter-type FS methods (MRMR FS and Chi-Square FS) were implemented in this study. MRMR FS was conducted using Matlab's fscmrmr-function that ranks the features according

to their relative importance in classification task using MRMR algorithm. In the used algorithm, the MI is used to measure the redundancy between the predictors and relevance between predictors and target variable as described in Chapter 4.6.1. Each feature is given a predictor importance score that reflects its importance, and the features are ranked based on this score in descending order. To make it possible to use both continuous and categorical features in the analysis, the algorithm discretizes continuous predictors by dividing them into 256 bins.

Chi-Square FS was performed using Matlab's fsschi2-function that constructs a univariate feature ranking for the features in the given data based on the results of Chi-Square tests. The feature importance score returned by the function equals to the logarithm of p-value of the Chi-Square test between a corresponding predictor and target variable. In cases where the p-value is smaller than eps(0), specifically $4.9407*10^{-324}$, the returned feature importance score is determined as infinite. To make use of mixed types of features (both continuous and categorical features), the continuous variables are again discretized into bins by the algorithm. By default, 10 bins are used. The default value for the bins was used because changing the number of bins was not found to have a considerable effect on the feature ranking.

The results of MRMR and Chi-Square FS are visualized in Figure 18. In the figure, the predictor importance scores of variables are plotted against their rankings in descending order (the predictor importance score of the most important feature is plotted first). The drop in the relative importance score indicates the confidence of FS. As it can be seen from the figure, the predictor importance scores decrease relatively fast in the beginning as the predictor rank gets lower. However, the relative drops decrease quickly, and after the rank of 30 the differences in relative importance cannot be considered notable.

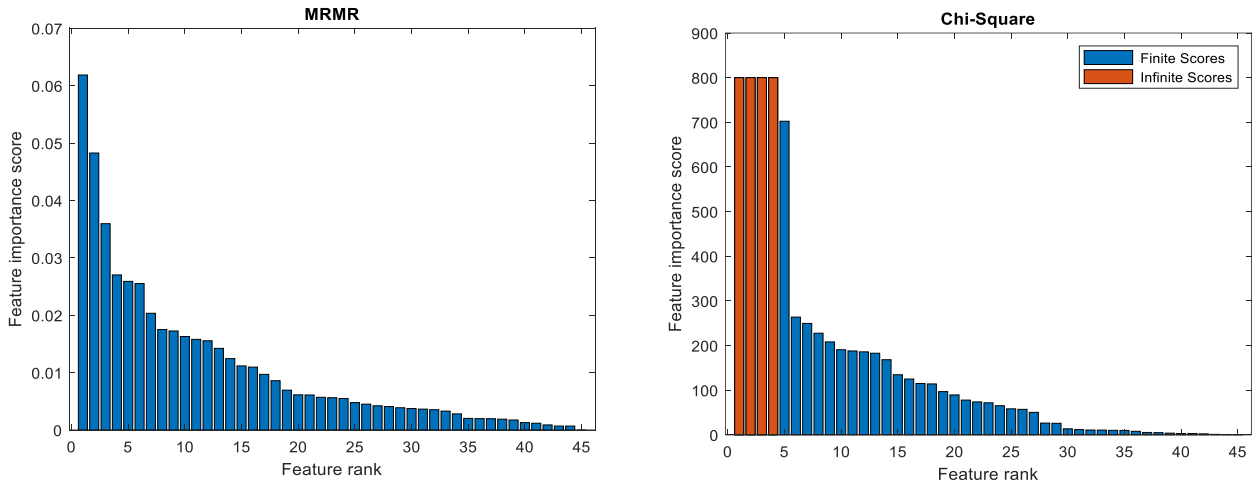


Figure 18. Visualization of filter-based FS results

The most important predictors ranked by different filter-type FS algorithms are listed in Table 8. The credit rating assigned by Bondora has the highest rank of features in the dataset according to both filter-type FS methods. In case of MRMR algorithm, it is followed by loan duration, free cash, education, and duration to first payment. According to the Chi-Square FS method, the next important features are the residency of the borrower, the maximum interest rate, language code and credit score assigned by the credit rating agency. It is noteworthy that the actual rankings of 4 most important features according to Chi-Square FS method are only directional (the superiority between these features cannot be judged exactly) because the importance scores for all of them are determined as infinite by Matlab.

Table 8. The most important predictors in case of filter FS methods

Rank	MRMR FS	Chi-Square FS
1	Credit rating	Credit rating*
2	Loan duration	Country*
3	Free cash	Interest*
4	Education	Language code*
5	Duration to first payment	Credit score
6	Amount of previous repayments	Amount of previous repayments
7	Monthly payment day	Monthly payment day
8	Gender	Bids manual
9	Home ownership type	Number of previous loans
10	Credit score	Amount of previous loans

*There are multiple infinite feature importance scores, so the ranking is directional.

Figure 19 shows the learning curves of different classifiers for the validation data when features are added sequentially to the classification models according to the order proposed by the filter-based FS methods. The selected number of features in case of each classifier is marked in the figure with a dashed line. Furthermore, the final results are listed in Table 9. As discussed earlier, the final decision of used features was made based on the CV error.

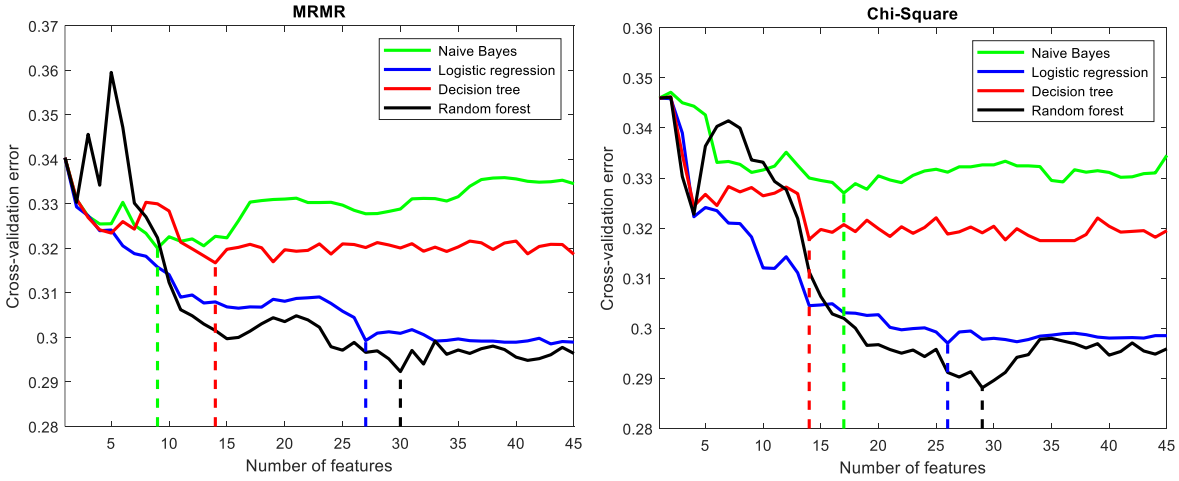


Figure 19. Learning curves of different classifiers (filter-type FS)

As it can be seen from the figure, in case of MRMR FS method the development of CV errors for the NB, LR and DT classifiers seems to be logical. The CV error curves are decreasing in the beginning when more features are added to the model: the added variables seem to be relevant for the classification task and improve the classification performance. After some point (as the feature ranking gets lower), the learning curves become stable or even turn downwards which indicate that the added features do not add useful information to the model. Instead, the noisiness of features can even decrease the classification performance of the models and hence increase the CV error. However, the learning curve of the RF classifier in case of MRMR FS differs from others: the CV error seems to turn upwards already after 2 added features. Nevertheless, the CV error starts to decrease again when 6th feature is added and keeps decreasing until the number of features is 30. For the RF classifier, 30 features were selected as the final feature set because at this point, the CV error is minimized.

The chosen number of features for the NB classifier was 9 because at this point, the CV error reaches its minimum. For the DT model, 14 features were selected to final feature set. Even if the CV error for DT would be slightly lower when 19 features were selected, the very slight improvement in accuracy is not enough to compensate the increasing model complexity. For the LR classifier, the selected number of features was 27. Again, the CV error obtained at this point for LR is not the absolute minimum of the CV error curve but the small improvement in accuracy cannot be considered significant enough to compensate the effects of increasing number of features after that point. This decision is supported by the fact that the importance scores provided by the algorithm are very low when the number of features exceeds 30.

Table 9. The final results of filter-based FS

FS method / classifier	NB	LR	DT	RF
MRMR	**9**	**27**	**14**	**30**
Chi-Square	**17**	**26**	**14**	**29**

The learning curves of different classifiers with Chi-Square based FS are relatively similar than in the case of MRMR FS. For the NB classifier, the selected number of features was 17 because CV error curve reaches its minimum at this point. For the DT model, 14 features were selected. Even if the CV error is again slightly lower after that point with few bigger feature sets, the small improvement in accuracy is not enough to recompense the increasing model complexity. For the LR model, the selected number of features was 26 and for the RF model, 29 features were selected.

6.4.2 Sequential forward selection

A wrapper-type SFS algorithm was used with each classifier to select the optimal feature subset. The 5-fold CV error was used as the evaluation criterion and to save the computational time, the SFS algorithm was run in parallel using Matlab's parallel computing toolbox. Because the execution time of the SFS increases considerably when the number of features increases, the label encoding was used for encoding of categorical variables in case of SFS algorithm. To begin with, Matlab's default option of the algorithm was used which stops the search when the first local optimum of the objective function is found. The results of SFS with different classifiers using the default option are listed in Table 10.

Table 10. The most important features proposed by SFS algorithm with default options

Feature rank	Naïve Bayes	Logistic regression	Decision tree	Random forest
1	Credit rating	Credit rating	Credit rating	Credit rating
2	Country	Language	Language	Language
3	Loan duration	Loan duration	Loan duration	Loan duration
4	Education	Country	Verification type	Country
Proposed features	8	14	6	4

As it can be seen from the table, the number of proposed features is relatively low: the algorithm proposes 8 features for the NB classifier, 6 features for the DT model and 4 features for the RF classifier. However, with the LR classifier, the proposed number of features is 14. It is also noteworthy that the proposed feature subsets are relatively similar: credit rating assigned by Bondora is the first proposed feature and loan duration is also among 4 first proposed features in all cases. Also, the language and country of the borrower seem to be among the most important features according to the SFS algorithm.

Because the SFS algorithm is prone to get stuck in a local optimum the objective function (and therefore the optimality of the feature subset cannot be guaranteed), the algorithm was also run so that it was forced to include all the features sequentially in the feature subset. The 5-fold CV errors with different classifiers are visualized in Figure 20 with all possible numbers of predictors. The CV error measured on the y-axis is the 5-fold CV error divided by the number of instances in validation sample. The selected number of features in case of each classifier is marked in the figure with dashed line. Furthermore, the final results of SFS are represented in Table 11.

For the NB classifier, the algorithm proposes 8 features to be used as a final feature subset with default settings. The decision is also supported by the visual analysis: when features are added to the feature subset, the value of the objective function decreases until it reaches its minimum when the number of features is 8. After that point, the CV error increases or remains the same when more features are added to the feature subset.

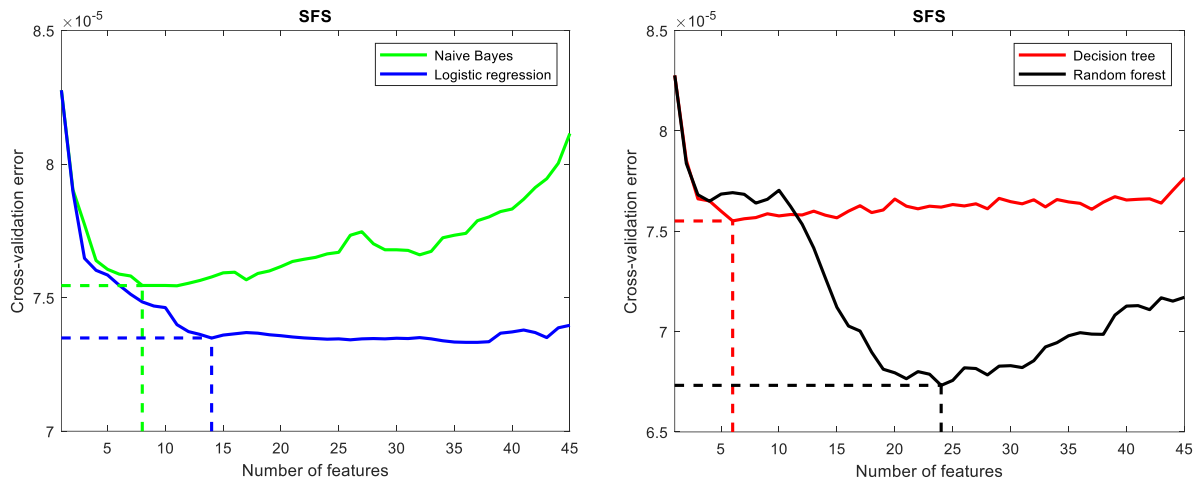


Figure 20. The results of SFS with different classifiers

However, the decision is not so clear for the LR classifier. As shown in Figure 20, the value of the objective function seems to increase clearly when the first features are included and keeps decreasing until it reaches a local optimum when the number of predictors is 14. However, instead of starting to increase after this point, the CV error stays somewhat stable until all the predictors are included. The minimum of the objective function is reached when the number of predictors is 37. This result highlights the drawback of the SFS algorithm being prone to get stuck in local optima. However, the slight decrease of CV error obtained by adding more features to the model does not compensate the increase in model complexity and therefore 14 features is still considered as optimal.

Table 11. The final results of sequential forward selection

Model	Naïve Bayes	Logistic regression	Decision tree	Random forest
Number of features	8	14	6	24

In case of DT classifier, the SFS algorithm with default options proposed 6 features to be used in the final classification. This is also chosen to be the final number of features after the graphical analysis. The result with the RF classifier is more complicated: the objective function improves at the beginning and has a local optimum when the number of predictors is 4. After that, the CV error stays somewhat stable until the number of predictors is 10. Afterwards, the CV error starts to decrease again. The minimum CV error is obtained at the point where the number of predictors is 27. The error at this point is markedly lower than in the case of 4 predictors, so 27 is considered to be the optimal number of features in case of RF classifier.

6.4.3 Learning-model based feature ranking

To also test the performance of LMBFR method, the feature importance scores for the DT and RF models were estimated and they were used for FS. Matlab's predictor importance estimation was used to estimate the relative importance of features. In case of DT, the importance of each feature was estimated by summing up the changes in the risk associated with the splits on corresponding feature and dividing this sum by the number of branch nodes. In case of RF, the predictor importance scores were estimated using the permutation importance measure which is described in more detail in Chapter 4.6.4.

It is noteworthy that the standard CART algorithm tends to fail to consider the interactions between features and also tends to select the features with many levels more often compared to the features with fewer levels (usually, CART prefers the continuous variables to categorical ones). Due to that, the predictor selection was done for both classification models with *interaction test* proposed by Loh (2002). This approach takes the interactions between the features better into account and considers the heterogeneous of variables and, therefore, offers more reliable estimates of relative feature importance than the standard CART algorithm.

10 most important features of each model according to the feature importance estimates are represented in Figure 21. As the figure shows, the most important features are again relatively similar: 6 of 10 most important features are the same across the different models. In the DT model, language and country of the borrower are 2 most important features, followed by the credit rating assigned by Bondora, the credit score assigned by third party and the maximum interest rate. In case of RF model, the loan duration and the maximum interest rate are 2 most important features, followed by the credit rating assigned by Bondora, the language and the amount of previous repayments.

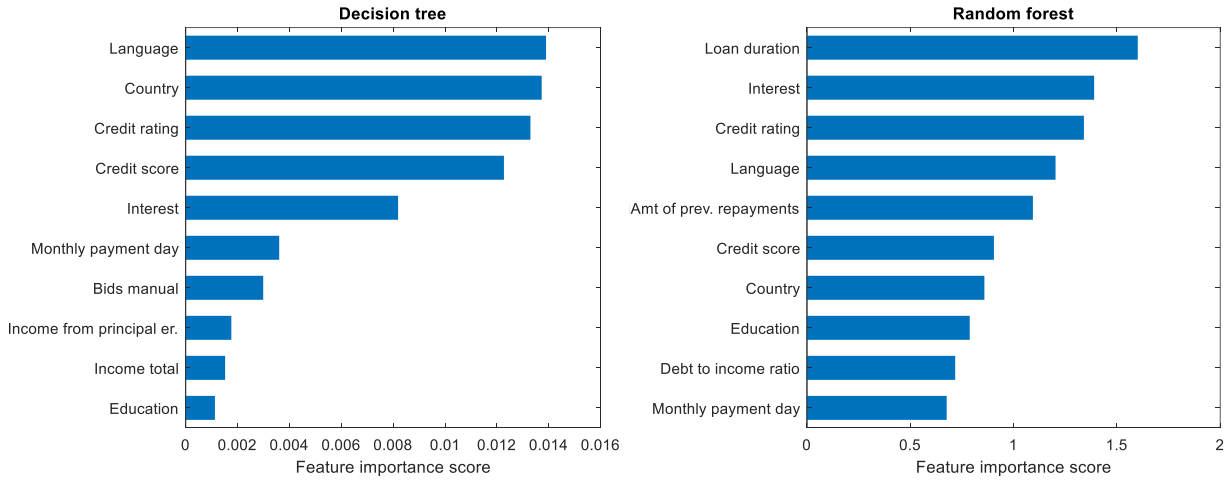


Figure 21. Feature importance scores of different classifiers

The learning curves of different classifiers are shown in Figure 22, representing the development of 5-fold CV error when the features are added sequentially to the model according to their feature rankings. The selected number of features is again marked in the figure with a dashed line for each classifier.

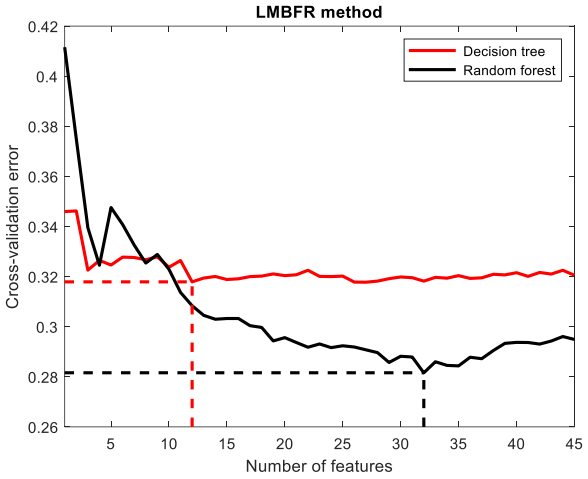


Figure 22. Learning curves of different classifiers (LMBFR method)

The number of selected features for the DT classifier was 12 because after this point, considerable improvements in CV performance were not achieved by adding new features. For the RF model, 32 features were selected because at this point, the minimum of the CV error curve was reached.

6.5 Choosing the hyperparameters and model training

After the data preprocessing and feature selection phases, the hyperparameters of the models were tuned and the final model specifications were selected. In this chapter, the hyperparameter optimization and model training phases are described.

6.5.1 Naïve Bayes

Matlab's fitcnb-function was used to train different combinations of the NB classifier and FS methods. The NB classifier is a simple classification model which has very few optimizable hyperparameters. Also, it is noteworthy that using the smoothing functions would have increased the computational time considerably and they were not found to have a notable effect on the model performance. Therefore, Matlab's default options were used in training in case of NB classifier. The used data distributions were Gaussian (normal) distribution for continuous variables and multivariate normal distribution for categorical variables.

Appendix 11 shows the in-sample and 5-fold CV errors for the different combinations of FS methods and the NB classifier. The CV and in-sample errors are close to each other with all the combinations of the NB classifier and different FS methods which indicates that notable overfitting problem is not present. Therefore, the default settings for the hyperparameters were considered reasonable in case of NB model.

6.5.2 Logistic regression

The LR classifier was trained with Matlab's fitglm-function using one by one different feature subsets proposed by different FS methods. Because the hyperparameter optimization is not the main focus of this study and the limitations of the used software package constrained the possibilities for hyperparameter optimization, the hyperparameters of the LR models were not systematically optimized but Matlab's default parameters were used in the model training.

The *modelspec* argument was set to "linear" which means that the fitted model includes an intercept and a linear term for each predictor. The *distribution* argument determining the distribution of the target variable was set to "binomial". Hence, the fitted models are generalized linear models with binomial response and logit link function. Regularization was not used because notable overfitting problem was not observed, and the number of observations in the dataset is considerably larger than the number of used features. Furthermore, the used fitting function does not support the use of regularization.

5-fold CV accuracy and the in-sample classification accuracy on the training data for all the LR model combinations are represented in Appendix 12. As it can be seen, the 5-fold CV accuracy and the in-sample accuracy of different models are relatively close to each other, indicating that the LR models do not overfit the training data. Therefore, the model with default parameters was considered sufficient for the purposes of this study.

6.5.3 Decision tree

In case of decision tree, Matlab's fitctree-function was used to train the classifiers with different subsets of features. The effect of hyperparameters can be considered relatively important in case of DT model because it is relatively prone to overfit the training data (Kotsiantis 2007; Mantovani et al. 2018), and therefore the hyperparameters of the DT models were optimized using Bayesian optimization in Matlab. It is noteworthy that the optimal hyperparameter values vary according to the used feature subset and hence the hyperparameter tuning was conducted for each model separately. The optimized hyperparameters were *minimum leaf size, maximum number of splits* and *split criterion.*

First two hyperparameters affect the complexity of the grown tree. The minimum leaf size defines the number of samples required in each leaf of the tree and the maximum number of splits defines the maximum number of branch nodes in the tree. The split criterion has two possible values for two-class classification problems, namely “gdi” and “deviance”. The first option determines Gini’s diversity index to be used as a split criterion and the second option exploits the maximum deviance reduction (also referred to as the cross entropy) as a split criterion.

Matlab’s default values for different hyperparameters are:

- Minimum leaf size = 1
- Maximum number of splits = N-1, where N = number of observations
- Split criterion: gdi (Gini’s diversity index)

The optimized hyperparameter settings for the different models are represented in Table 12. As it can be seen from the table, the optimal minimum leaf size is higher than the default value in almost all the cases. This can be explained by the fact that by using the default minimum leaf size (1), the DT classifier tends to grow very complex tree. The more complex tree is more prone to overfit the training data, and overfitting decreases the generalization performance of the model. When the minimum leaf size is 1 (an extreme case), only 1 observation is required in each leaf of the tree which typically leads to a large number of splits, complicating the grown tree. The maximum number of splits in the tree is also much smaller in all the cases compared to the default value. The smaller number of splits decreases the complexity of the tree which again helps to avoid the overfitting problems.

Table 12. Optimized hyperparameters of DT

Model	Min. leaf size	Max. number of splits	Split criterion
DT (No FS)	10	14	deviance
DT + MRMR	4	39	gdi
DT + Chi-Square	2	42	deviance
DT + SFS	182	107	gdi
DT + LMBFR	2	11	gdi

The same conclusions can be drawn by comparing the in-sample classification error to 5-fold CV error with different hyperparameter settings. These metrics are shown in Appendix 13 for the DT models with both default and optimized hyperparameters. In general, the in-sample classification error with default hyperparameter values is much lower than the 5-fold CV error which indicates overfitting.

In contrast, the in-sample and CV errors are near to each other when the classification is conducted with the optimized hyperparameters. The hyperparameter optimization also decreases the CV error notably in all the cases. This indicates that the hyperparameter optimization can reduce the overfitting problem of the model and improves the out-of-sample performance of the classifier.

6.5.4 Random forest

The RF models were trained using Matlab's fitcensemble-function. The function uses by default the random forest algorithm proposed by Breiman (2001) to select the predictors randomly at each split when used to fit bagging ensembles. The hyperparameters of RF models were optimized using Bayesian hyperparameter optimization. The optimized hyperparameters included *number of learning cycles, number of variables to sample, maximum number of splits, minimum leaf size* and *split criterion.* The number of learning cycles defines the number of grown trees in the forest and the number of variables to sample defines the number of variables which are chosen randomly to construct each tree (Probst, Wright, Boulesteix 2019). The functions of other hyperparameters were explained earlier in the previous chapter when discussing the hyperparameter optimization of DT models.

Matlab's default values for different hyperparameters are:

- Number of learning cycles = 100
- Number of variables to sample = Square root of the number of predictors
- Minimum leaf size = 1
- Maximum number of splits = N-1, where N = number of observations
- Split criterion: gdi (Gini's diversity index)

Table 13 shows the hyperparameter values found with Bayesian optimization for each combination of the RF classifier and feature selection methods. In the optimization, the range of 1 to 500 was assigned for the number of learning cycles and the maximum number of splits. For the RF classifier, the optimization results are very different compared to the results for DT. The optimal minimum leaf size seems to be near the default value (1) in all the cases except from the LMBFR method. The maximum number of splits is also higher than in the case of DT. This can be explained by the different characteristics of the RF classifier: averaging the results across the trees grown in a forest (ensemble of decision trees) helps to avoid overfitting. Optimal numbers of learning cycles (grown trees) for the models are between 114 and 459.

Table 13. Optimized hyperparameters of RF

Model	No. of learning cycles	No. of variables to sample	Min. leaf size	Max. number of splits	Split criterion
RF (No FS)	342	4	1	499	deviance
RF + MRMR	389	7	2	487	deviance
RF + Chi-Square	450	14	1	497	gdi
RF + SFS	459	13	6	496	deviance
RF + LMBFR	114	30	25	495	gdi

The 5-fold CV errors of different RF models are represented in Appendix 14 with default parameters and with the hyperparameter settings obtained using Bayesian optimization. In fact, it seems that based on CV error comparison, the Bayesian search fails to find the optimal hyperparameters for the RF models, and the CV error obtained using the optimized hyperparameters is higher in all the cases.

However, it is noteworthy that the differences between CV errors before and after Bayesian optimization are relatively small regarding some of the models. Therefore, the final classification results are still investigated using both default hyperparameters and the ones received from Bayesian optimization.

6.6 Evaluation of different methods

Following the established procedures of previous related studies (see Chapter 5 for details), the classification performance and the model complexity are used in the comparison of different FS methods in this thesis. The classification performance obtained with the feature subset proposed by different FS algorithms is compared both to the classification performance obtained using the model without FS and to the classification results with different FS methods. The model complexity is evaluated by comparing the number of proposed features in case of different FS methods.

Classification accuracy, sensitivity and specificity (see Chapter 3.4 for details) are used as classification performance measures. Because the purpose of this study is to predict the loan defaults, more weight is given for the true positive rate (sensitivity) than for the true negative rate (specificity) in the analysis. In other words, classifying the default loans (positive instances) into non-default (negative) class is considered more harmful than classifying the non-default loans into default class. In addition to the confusion matrix -based measures, the AUC metric is used to measure the final classification performance.

The statistical significance of results is tested with McNemar test which has been suggested by Dietterich (1997b) for comparing supervised classification learning algorithms in cases

where the performance is evaluated using holdout validation. The two-sided mid-p test version (Matlab's default) of McNemar test was used to test for the statistical significance of differences between accuracies of different models. The accuracy of each classification model without FS was used as a benchmark to investigate the statistical significance of the effect of FS on model accuracy.

6.6.1 Classification performance

The final classification performance of the NB classifier with different FS methods is represented in Table 14. The accuracies obtained using different FS methods are all slightly higher than the accuracy obtained with the full feature set. The improvement in accuracy is the biggest with the SFS method which helps to increase the accuracy by 1.34% (from 0.667 to 0.676). Also, the AUC scores with all the FS methods are slightly higher compared to the benchmark model. The sensitivity (true positive rate) of the model also increases with all the FS methods, and especially with the MRMR method, the increase in sensitivity is high (23.64% relative change). However, the specificity of the NB model decreases with the MRMR FS method by 25.82%, hence distorting the overall accuracy. According to McNemar test, the improvement in accuracy is statistically significant at 5% risk rate only in case of SFS method.

Table 14. Final classification results with NB classifier

Model	Accuracy	Sensitivity	Specificity	AUC	p-value
NB without FS	0.667	0.640	0.701	0.705	-
NB + MRMR	0.672	0.791	0.520	0.710	0.266
NB + Chi-Square	0.671	0.650	0.698	0.713	0.147
NB + SFS	0.676	0.710	0.632	0.730	0.029**

The p-values are from McNemar test which is used to test for statistical significance of differences in model accuracy.
Significance codes: *** p-value significant at α = 0.01, ** p-value significant at α = 0.05, * p-value significant at α = 0.10

In case of different LR models, the final performance results are shown in Table 15. The full feature set and the feature subsets proposed by different FS algorithms lead approximately to the same classification accuracy (0.697–0.700). The full feature set provides the best classification results measured by sensitivity (0.814) and AUC score (0.755). However, the specificity obtained using the full feature set is the worst across all the models (0.552). This means that the model with full feature set fails to classify the non-default instances to correct (negative) class. Using each of the FS methods leads to at least competitive classification performance compared to the results with full feature set. The differences in the classification accuracy are not statistically significant and the differences in AUC score between models does not seem to be considerable either.

Table 15. Final classification results with LR classifier

Model	Accuracy	Sensitivity	Specificity	AUC	p-value
LR without FS	0.699	0.814	0.552	0.755	-
LR + MRMR	0.699	0.780	0.596	0.751	1.000
LR + Chi-Square	0.700	0.815	0.553	0.753	0.741
LR + SFS	0.697	0.770	0.604	0.747	0.539

The p-values are from McNemar test which is used to test for statistical significance of differences in model accuracy.
Significance codes: *** p-value significant at α = 0.01, ** p-value significant at α = 0.05, * p-value significant at α = 0.10

The final classification results of DT models are represented in Table 16 before and after the hyperparameter tuning. With the default hyperparameters, the classification performance is improved using every FS method except from the LMBFR method. The improvement in classification performance is the biggest with the SFS algorithm which helps to improve accuracy by 7.74% (by 0.047 in absolute value). Simultaneously, the AUC score increases from 0.600 to 0.692 (15.31%) and the sensitivity increases from 0.656 to 0.708 (8.05%). According to McNemar test, the improvement in accuracy with the SFS method is also statistically significant at 1% significant level. In case of other methods, the accuracy improvements are not statistically significant. However, it is noteworthy that the improvements in AUC are considerable also in case of MRMR and Chi-Square FS methods (6.41% and 7.14%, respectively).

Table 16. Final classification results with DT classifier

	Before hyperparameter tuning					After hyperparameter tuning				
Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**p-value**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**p-value**
DT without FS	0.610	0.656	0.551	0.600	-	0.687	0.797	0.546	0.712	-
DT + MRMR	0.615	0.643	0.580	0.639	0.408	0.686	0.797	0.544	0.729	0.837
DT + Chi-Square	0.620	0.663	0.565	0.643	0.104	0.683	0.793	0.542	0.726	0.199
DT + SFS	0.657	0.708	0.591	0.692	1.36E-13***	0.685	0.776	0.569	0.728	0.733
DT + LMBFR	0.604	0.655	0.537	0.618	0.371	0.686	0.803	0.535	0.701	0.295

The p-values are from McNemar test which is used to test for statistical significance of differences in model accuracy.
Significance codes: *** p-value significant at α = 0.01, ** p-value significant at α = 0.05, * p-value significant at α = 0.10

When the hyperparameter optimization is conducted, the improvements in the classification performance obtained using FS are not statistically significant anymore. The accuracies with all the FS methods after the hyperparameter tuning (0.683-0.686) are even slightly worse than without FS (0.687). However, the AUC scores are slightly higher with all the FS methods except from LMBFR method. It is noteworthy that all the performance measures except from the specificity are improved by performing the hyperparameter optimization in case of all the feature subsets (and the specificity also stays at a comparable level compared to situation before optimization). The results can be explained by the characteristics of the DT classifier: with the default hyperparameters, the DT model tends to grow a very complex tree which is prone to overfit the training data so that the test performance is deteriorated. When the FS methods are applied and the number of features is reduced, the generalization ability of the model improves

compared to the performance before FS. However, when the overfitting problem is handled by tuning the hyperparameters, the effect of proper FS seems not to be significant anymore.

Finally, the classification results of RF models both before and after the hyperparameter tuning are represented in Table 17. Before hyperparameter tuning, all the FS methods except from the Chi-Square method manage to provide better classification accuracy than the model without FS. Even with the Chi-Square FS method, the accuracy of the model is the same (expressed to three decimal places) as the accuracy of the model without FS. The best accuracy is again obtained using the SFS method, which manages to improve the accuracy by 4.49% (from 0.708 to 0.739). Also, the LMBFR seems to improve the accuracy considerably (from 0.708 to 0.718). Both improvements in accuracy are also statistically significant at 1% significance level. Furthermore, the AUC scores are at least slightly higher with all the FS methods compared to the model without FS. The sensitivity is also improved considerably by using the SFS method (by 3.49%), and the specificity also improves with all the FS methods compared to the model without FS.

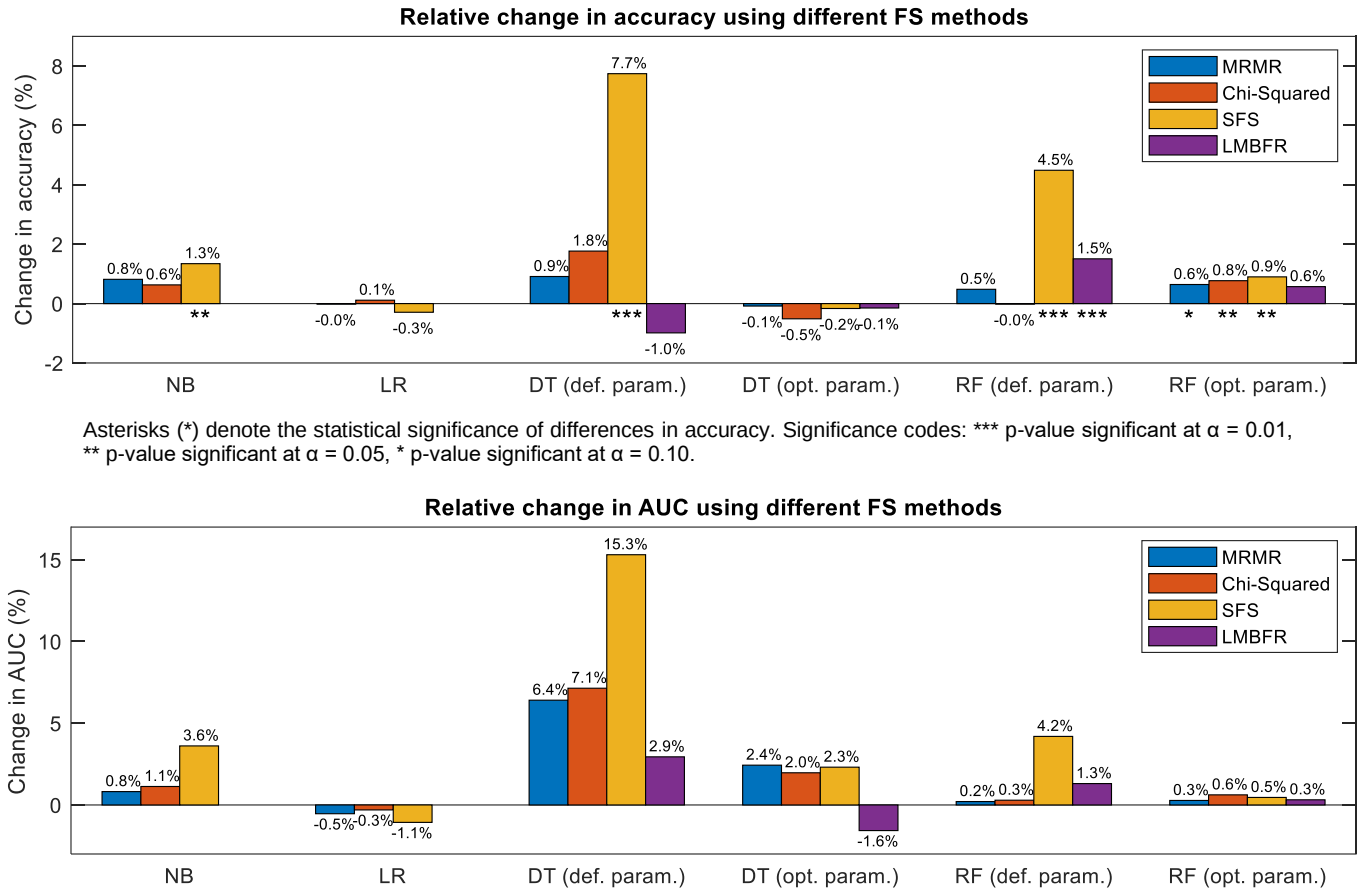
Table 17. Final classification results with RF classifier

	Before hyperparameter tuning					After hyperparameter tuning				
Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**p-value**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**	**p-value**
RF without FS	0.708	0.788	0.605	0.766	-	0.700	0.764	0.619	0.761	-
RF + MRMR	0.711	0.784	0.618	0.767	0.278	0.705	0.781	0.607	0.763	0.064*
RF + Chi-Square	0.708	0.783	0.611	0.768	0.972	0.706	0.789	0.599	0.766	0.049**
RF + SFS	0.739	0.816	0.642	0.798	2.32E-17***	0.707	0.797	0.592	0.764	0.050**
RF + LMBFR	0.718	0.796	0.619	0.776	0.001***	0.704	0.791	0.594	0.763	0.200

The p-values are from McNemar test which is used to test for statistical significance of differences in model accuracy.
Significance codes: *** p-value significant at α = 0.01, ** p-value significant at α = 0.05, * p-value significant at α = 0.10.

As stated earlier, the Bayesian optimization conducted for hyperparameters was found not to improve the CV performance of the RF models. This is the case also regarding the final classification results. Actually, the classification performance measured by accuracy and AUC even decreases after the hyperparameter tuning compared to the results with default hyperparameters in all cases. Also, the sensitivity and specificity of the models seem to worsen in most cases. The classification performance is considerably lower than the performance before hyperparameter tuning in case of SFS and LMBFR methods.

However, when the models after the hyperparameter tuning are compared to each other, all the FS methods except from the LMBFR seem to improve the classification accuracy statistically significantly at least at 10% significance level compared to the model without FS. Also, the AUC score and sensitivity of each model improve at least slightly compared to the benchmark model.



Asterisks (*) denote the statistical significance of differences in accuracy. Significance codes: *** p-value significant at α = 0.01, ** p-value significant at α = 0.05, * p-value significant at α = 0.10.

Figure 23. Visualization of changes in accuracy and AUC using different FS methods

The relative changes in accuracy and AUC values using different FS methods are further visualized in Figure 23 and some conclusions can be drawn from the graphical analysis. Firstly, the SFS method seems to be the most accurate from the FS methods. It seems to improve both the accuracy and AUC metrics in most cases. Secondly, it is noteworthy that FS seems not to have a significant impact on performance of the LR classifier. Thirdly, visual analysis highlights the finding made earlier that the effect of FS decreases when the hyperparameter optimization is conducted. This appears to be logical since hyperparameter optimization and FS are frequently used to reach the same goals: both are frequently used to avoid the overfitting problems of the prediction models.

6.6.2 The number of selected features (model complexity)

The performance of used FS methods was also investigated by comparing the number of selected features (model complexity) with different classification models. The number of features used with different combinations of FS methods and classification models are represented in Figure 24. The y-axis of the chart is limited to range from 0 to 45, which is the maximum number of features. As the figure shows, all the FS methods can reduce the number of features considerably. As discussed earlier, the classification accuracy did not decrease statistically

significantly in any case, so all the FS methods can be stated to reduce the complexity of the models markedly without significantly deteriorating the classification performance.

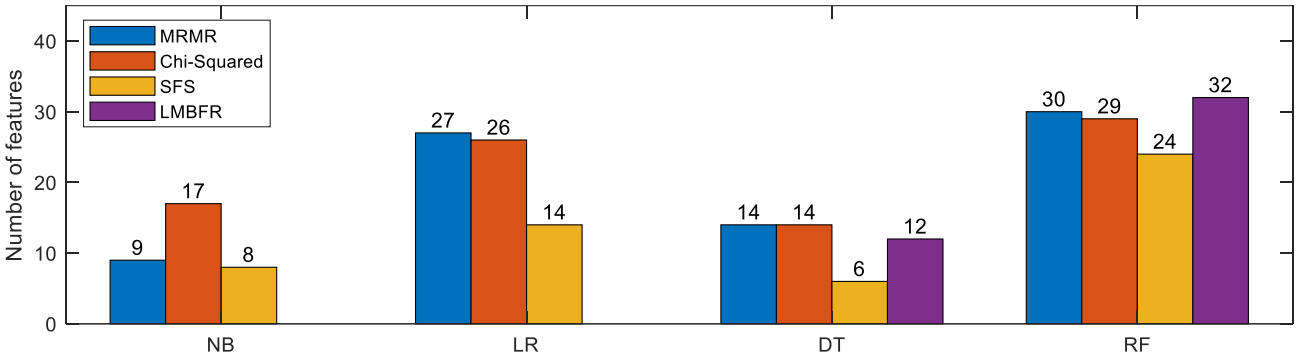


Figure 24. The number of selected features using different FS methods

The results show that the SFS algorithm can reduce the number of features more than other FS methods in case of all the classifiers. It is worth noting that at the same time, the improvements in classification performance are the most considerable in case of SFS algorithm which highlight the superior performance of the SFS method compared to other FS methods. With other FS methods, the selected number of features are somewhat similar for the certain classifiers. However, in case of Naïve Bayes classifier, the selected number of features with Chi-Square FS is higher than with other methods. It is noteworthy that the number of features seem to vary markedly between different classifiers: the NB and DT classifiers seem to reach the best classification accuracy with considerably lower number of features than the LR and RF classifiers in most cases. This can be explained by the different characteristics of tested classification models.

6.7 Determinants of default

As described earlier, the tree-based classification models offer an intuitive way to estimate the relative importance of features used in the models. These estimates (reported and discussed in more detail in Chapter 6.4.3 when conducting LMBFR FS) are used to identify the most important features in predicting the default in Bondora dataset. The statistical dependence analysis reported in Chapter 6.2.1 is also used as a supportive tool to examine the most important default determinants. However, it is worth noting that the used statistical measures (the point-biserial correlation coefficients and Chi-Square test statistics) provide only directional insight regarding the feature importance because they fail to consider the interrelationships between variables.

10 most important features according to the estimated relative feature importance scores of DT and RF classifiers are represented in Table 18. Besides that, 5 most important categorical

and continuous features according to the statistical dependence analysis (point-biserial correlation and Chi-Square tests) are listed in the same table.

Table 18. The most important determinants of default

Rank/ Method	Decision tree feature importance	Random forest feature importance	Rank/ Method	Point-biserial correlations and Chi-Square tests	
1	**Language**	**Loan duration**	**1**	**Credit score**	**Continuous features**
2	**Country**	**Interest rate**	**2**	**Interest rate**	
3	**Credit rating**	**Credit rating**	**3**	**Amt of previous repayments**	
4	**Credit score**	**Language**	**4**	Number of previous loans	
5	**Interest rate**	**Amt of previous repayments**	**5**	Amount of previous loans	
6	**Monthly payment day**	**Credit score**	**1**	**Credit rating**	**Categorical features**
7	Bids manual	**Country**	**2**	**Country**	
8	Inc. from principal employer	**Education**	**3**	**Monthly payment day**	
9	Income total	Debt to income ratio	**4**	**Loan duration**	
10	**Education**	Applied amount	**5**	New credit customer	

The bolded features are among the most important determinants according to at least two approaches.

As it can be seen from the table, the most important features according to all the used methods are relatively similar. Altogether, 4 features (credit score assigned by third party credit rating agency, interest rate, credit rating assigned by Bondora and the country of the borrower) are ranked to be among the most important features according to all the used methods. Furthermore, 5 other variables (language, amount of previous repayments, education, loan duration and monthly payment day) are ranked to be among the most important features according to at least two of the used approaches.

6.8 Analysis and discussion of the results

In this chapter, the results of the performance comparison are analyzed and discussed, and the results are also compared to the findings of previous related studies. Furthermore, the second and third research questions of the thesis are answered. As stated in Chapter 6.4, the selection procedure used in this study to select the final number of features makes also the MRMR, Chi-Square and LMBFR methods technically wrappers. Thus, it is noteworthy that the differences in model performance between different FS methods actually indicate the effect of variations in used heuristics to construct the feature ranking in the first place (the MRMR-based ranking, Chi-Square test based ranking and learning-model based feature ranking).

6.8.1 Model performance

Table 19 represents the overall comparison of the results across all the tested FS and classification models. The table highlights the findings of model performance made earlier: the SFS method seems to provide the best performance from different FS methods. It manages to

improve the classification accuracy statistically significantly (at least at 5% significance level) in case of NB classifier (1.34%), DT classifier with default hyperparameters (7.74%) and RF classifier with both default and optimized hyperparameters (4.49% and 0.90%, respectively). Also, the improvements in AUC metric obtained using SFS method in case of NB classifier (3.61%), DT classifier with both default and optimized parameters (15.31% and 2.31%, respectively) and RF classifier with default parameters (4.19%) can be considered notable.

Table 19. The comparison of results across all the tested models

Classifier / FS method	No FS	MRMR	Chi-Square	SFS	LMBFR
Naïve Bayes	(a) 0.667	(a) 0.672 (0.82%)	(a) 0.671 (0.63%)	**(a) 0.676 (1.34%)****	(a) -
	(b) 0.705	(b) 0.710 (0.82%)	(b) 0.713 (1.13%)	**(b) 0.730 (3.61%)**	(b) -
	(c) 0.640	**(c) 0.791 (23.64%)**	(c) 0.650 (1.52%)	**(c) 0.710 (10.94%)**	(c) -
	(d) 0.701	**(d) 0.520 (-25.82%)**	(d) 0.698 (-0.41%)	**(d) 0.632 (-9.85%)**	(d) -
	(e) 45	(e) 9 (-80.00%)	(e) 17 (-62.22%)	**(e) 8 (-82.22%)**	(e) -
Logistic regression	(a) 0.699	(a) 0.699 (0.00%)	**(a) 0.700 (0.11%)**	(a) 0.697 (-0.29%)	(a) -
	(b) 0.755	(b) 0.751 (-0.54%)	(b) 0.753 (-0.30%)	(b) 0.747 (-1.07%)	(b) -
	(c) 0.814	(c) 0.780 (-4.25%)	**(c) 0.815 (0.10%)**	(c) 0.770 (-5.44%)	(c) -
	(d) 0.552	(d) 0.596 (8.01%)	(d) 0.553 (0.14%)	**(d) 0.604 (9.41%)**	(d) -
	(e) 45	(e) 27 (-40.00%)	(e) 26 (-42.22%)	**(e) 14 (-68.89%)**	(e) -
Decision tree (default parameters)	(a) 0.610	(a) 0.615 (0.91%)	(a) 0.620 (1.77%)	**(a) 0.657 (7.74%)*****	(a) 0.604 (-0.99%)
	(b) 0.600	**(b) 0.639 (6.41%)**	**(b) 0.643 (7.14%)**	**(b) 0.692 (15.31%)**	**(b) 0.618 (2.94%)**
	(c) 0.656	(c) 0.643 (-1.94%)	(c) 0.663 (1.20%)	**(c) 0.708 (8.05%)**	(c) 0.655 (-0.03%)
	(d) 0.551	(d) 0.580 (5.26%)	(d) 0.565 (2.63%)	**(d) 0.591 (7.27%)**	(d) 0.537 (-2.44%)
	(e) 45	(e) 14 (-68.89%)	(e) 14 (-68.89%)	**(e) 6 (-86.67%)**	(e) 12 (-73.33%)
Decision tree (optimized parameters)	**(a) 0.687**	(a) 0.686 (-0.08%)	(a) 0.683 (-0.51%)	(a) 0.685 (-0.17%)	(a) 0.686 (-0.15%)
	(b) 0.712	**(b) 0.729 (2.43%)**	**(b) 0.726 (1.97%)**	**(b) 0.728 (2.31%)**	(b) 0.701 (-1.57%)
	(c) 0.797	(c) 0.797 (0.08%)	(c) 0.793 (-0.43%)	(c) 0.776 (-2.51%)	**(c) 0.803 (0.81%)**
	(d) 0.546	(d) 0.544 (-0.38%)	(d) 0.542 (-0.66%)	**(d) 0.569 (4.21%)**	(d) 0.535 (-1.94%)
	(e) 45	(e) 14 (-68.89%)	(e) 14 (-68.89%)	**(e) 6 (-86.67%)**	(e) 12 (-73.33%)
Random forest (default parameters)	(a) 0.708	(a) 0.711 (0.48%)	(a) 0.708 (-0.02%)	**(a) 0.739 (4.49%)*****	**(a) 0.718 (1.51%)*****
	(b) 0.766	(b) 0.767 (0.21%)	(b) 0.768 (0.29%)	**(b) 0.798 (4.19%)**	(b) 0.776 (1.31%)
	(c) 0.788	(c) 0.784 (-0.51%)	(c) 0.783 (-0.59%)	**(c) 0.816 (3.49%)**	(c) 0.796 (1.05%)
	(d) 0.605	(d) 0.618 (2.14%)	(d) 0.611 (0.94%)	**(d) 0.642 (6.15%)**	(d) 0.619 (2.26%)
	(e) 45	(e) 30 (-33.33%)	(e) 29 (-35.56%)	**(e) 24 (-46.67%)**	(e) 32 (-28.89%)
Random forest (optimized parameters)	(a) 0.700	**(a) 0.705 (0.64%)***	**(a) 0.706 (0.77%)****	**(a) 0.707 (0.90%)****	(a) 0.704 (0.57%)
	(b) 0.761	(b) 0.763 (0.28%)	**(b) 0.766 (0.62%)**	(b) 0.764 (0.46%)	(b) 0.763 (0.32%)
	(c) 0.764	(c) 0.781 (2.29%)	(c) 0.789 (3.35%)	**(c) 0.797 (4.31%)**	(c) 0.791 (3.52%)
	(d) 0.619	(d) 0.607 (-1.97%)	(d) 0.599 (-3.29%)	(d) 0.592 (-4.47%)	(d) 0.594 (-4.10%)
	(e) 45	(e) 30 (-33.33%)	(e) 29 (-35.56%)	**(e) 24 (-46.67%)**	(e) 32 (-28.89%)

(a) Accuracy, (b) AUC, (c) Sensitivity, (d) Specificity, (e) Number of variables of different models. Values in parentheses are relative changes compared to the model without FS. The best value of each performance measure with each classification model (with both default and optimized hyperparameters) is underlined and bolded. Other most considerable values are bolded. Asterisks (*) denote the statistical significance of differences in accuracy. Statistical significance codes: *** p-value significant at $\alpha = 0.01$, ** p-value significant at $\alpha = 0.05$, * p-value significant at $\alpha = 0.10$.

However, some considerable improvements in classification performance were obtained also with other methods. Measured by AUC, the classification performance of the DT model (with default hyperparameters) improves considerably with the MRMR, Chi-Square and LMBFR methods (by 6.41%, 7.14% and 2.94%, respectively). With optimized hyperparameters, the improvement in AUC using the DT model can still be considered notable with the MRMR FS approach (2.43%) and Chi-Square FS method (1.97%). In addition, the improvement in accuracy of the RF model (with default hyperparameters) in combination with the LMBFR method

(1.51%) is statistically significant compared to the benchmark model. Furthermore, after the hyperparameter tuning, the accuracy improvements of the RF classifier obtained with both MRMR and Chi-Square FS methods (0.64% and 0.77%, respectively) are statistically significant at least at 10% significance level.

Overall, the obtained classification performance of the models can be considered potential. The overall accuracies of the best models of each classifier range from 0.676 to 0.739. When measured by AUC, the classification performance of the best-performing models of each classifier ranges from 0.729 to 0.798. The RF method was found to outperform the other used models in overall classification performance in general. The best RF model (with default hyperparameters and the SFS algorithm) provided the classification accuracy of 0.739 and AUC score of 0.798. The findings of RF classifier's good performance support for example the results of the study conducted by Malekipirbazari and Aksakalli (2015), in which the RF was found to perform best from the tested classification models in P2P lending context.

Based on the results of this study, the LR classification model is also suitable for the P2P lending default prediction and provides somewhat competitive performance with more sophisticated RF classifier. The best performing LR model in which the FS was conducted using the Chi-Square FS method resulted in the accuracy of 0.700 and AUC of 0.753. Similar findings have been made earlier by Teply and Polena (2020) who found in their study that the LR was competitive with more sophisticated classification models in P2P lending default prediction. In their study, the LR model even outperformed many more sophisticated classifiers (including for example the RF, SVM and NN based models).

The DT classifier provided the best classification accuracy of 0.687 with optimized hyperparameters and the full feature set but the accuracy in combination with each of the FS methods (with optimized hyperparameters) was on a competitive level (0.683–0.686). The best AUC across DT models after the hyperparameter optimization was obtained using the MRMR FS (0.729), but the AUC obtained using SFS method was also comparable (0.728).

Overall, NB classifier was found to perform the worst from the used classifiers according to the used classification performance metrics. The best classification accuracy in case of NB classifier was obtained in combination with SFS method (0.676). This model resulted in AUC of 0.730. The inefficiency of NB classifier can be seen logical because of model's simplicity and strong assumptions. However, it is noteworthy that the observed differences between NB and DT models' performance were not dramatic, and the best NB classifier even slightly outperformed the best DT model when measured by AUC.

The best sensitivity values obtained using each classifier range from 0.791 to 0.816. The best sensitivity (0.816) was obtained using the most accurate RF classifier and the most accurate LR model also provided the sensitivity of 0.815. This indicates that the models can somewhat efficiently identify the loans that are most likely to default: the models can efficiently predict the default loans to the default class. However, the specificity of the models was considerably lower than the sensitivity: the best specificity values of each classifier range from 0.569 to 0.701. This indicates that the good ability to identify the default instances comes at least partially at the expense of ability to identify the non-default instances – classification models are not as efficient in classifying the non-default loans to non-default class as they are in classifying the default loans to default class.

It is also worth noting that the characteristics of used classification models affect the performance of FS methods. The LR, DT and RF classifiers have their own in-built feature selection procedures, which help to assign more weight to the most relevant features. For example, the LR classifier weights the features automatically according to their predictive performance (Liu and Schumann 2005), and FS methods in combination with the LR model did not appear to provide improvements in classification performance in this study. However, even though the tree-based models also select automatically the most relevant features to be used in the classification tree structure (Liu and Schumann 2005), the performance of the DT and RF classifiers was still improved by using FS methods (especially in case of SFS method).

In contrast to other used classifiers, the NB model is a very simple classification algorithm that weights all the used features equally (Chang-Hwan, Gutierrez, Dejing 2011). Logically, in case of NB classifier, all the FS methods were found to provide at least slightly better classification performance compared to the results with the full feature set even though the improvement in accuracy was significant only in case of SFS method.

In addition, it is remarkable that the good performance of SFS method was obtained at the expense of computational efficiency – the SFS method was found to be computationally more expensive than the filter-type MRMR and Chi-Square methods and the embedded-type LMBFR method. However, it is worth noting that the computational time of the FS methods was not examined systematically in this study, and the closer investigation of the trade-off between classification performance and the computational time of different FS methods is left to the future research.

To conclude, it can be stated that especially the results of the best-performing classification models can be considered potential in P2P lending default prediction, but all the models are still far from perfect classification performance. This hints at the fact that in P2P lending, other

relevant aspects of both borrowers and loans that are not collected by the P2P lending platforms may also have a considerable effect on the probability of default.

6.8.2 Answering the research questions

The first research question and its three sub-questions were related to the previous research of the study subject. The research question was answered earlier after conducting the literature review (in Chapter 5.6) and replicating the answer here is not considered necessary.

The second research question was the main research question of this thesis and it was related to the performance of different FS methods. It was formed as follows:

"How do different feature selection methods perform compared to each other in P2P lending default prediction?"

To answer the research question, both the final classification performance and the number of features (model complexity) are considered. SFS model was able to reduce the number of features most considerably in case of all classifiers. It also managed to provide statistically significant improvement in classification performance with most of the tested classification models and provided the best classification performance of all the used models in combination with RF classifier. Therefore, the SFS method is stated to perform the best from the tested FS methods on the Bondora dataset.

The good performance of the SFS method can be at least partially explained by the fact that the SFS algorithm as a wrapper-type FS method incorporates the corresponding classifier in the FS process and exploits the classification accuracy as the evaluation criterion in feature subset evaluation. This typically leads to relatively high final classification accuracy. Another explanation can be that the SFS method can consider the interrelations between features better than the used filter and embedded FS methods (Kohavi and John 1997). The finding of superiority of wrapper-type FS method over other methods is in line with the previous research in consumer credit scoring area in which the wrapper-type FS methods have been found most efficient in terms of classification performance (Liu and Schumann 2005; Somol et al. 2005).

When comparing the performance of other methods, there are no clear winners. All the FS models except from the SFS method lead to the selection of about the same number of features for specific classifiers. Also, the differences in classification performance using other FS methods are not systematic across different classifiers. All these FS methods lead to the statistically significant improvement in classification accuracy only in case of one classifier.

Therefore, the SFS method is stated to perform the best across different FS methods but other methods are stated to perform somewhat equally compared to each other.

However, even though the systematic improvements in classification performance were not observed using other FS methods (except from the SFS method), all the tested FS methods can still be considered suitable in P2P lending default prediction. This is due to the fact that all the used FS methods can reduce the used number of features in classification models (model complexity) considerably while providing at least competitive classification accuracy (without significant reduction in overall classification accuracy). That is beneficial because the reduced complexity improves the interpretability and understandability of the models and helps to make the models less computationally expensive.

The third research question of the thesis was related to the most important predictors of P2P loan default. It was formed as follows:

> "What are the most important features in predicting the default in Bondora dataset?"

The research question can be answered based on the analysis conducted in Chapter 6.7. According to the results of the analysis, it can be said that there seems to be several important features in predicting the default in Bondora dataset. The credit rating determined for each loan by Bondora and the credit score assigned to each borrower by a third party credit rating agency were found to be among the most important determinants of default. These findings support the results of previous studies conducted using other P2P lending datasets: the credit ratings assigned by P2P platforms have appeared to be significant predictors of default (Emekter et al. 2015; Malekipirbazari and Aksakalli 2015; Serrano-Cinca et al. 2015). Also, the demographics of the borrower such as residency, language and education seem to be important determinants of default risk in Bondora data. These results are also in line with previous studies in P2P lending area (Byanjankar et al. 2015; Xia et al. 2017; Lin et al. 2017).

Furthermore, the loan characteristics, especially the maximum interest rate determined by the borrower and the duration of the loan strongly affect the default risk based on the results of this study. Similar findings have been earlier made in the P2P lending context for example by Jin and Zhu (2015), Chen et al. (2017), and Xia et al. (2017). In addition, earlier credit and payment history seem to have an effect on default probability in Bondora data. These results also support the results of previous research conducted in P2P lending area (Polena and Regner 2018; Serrano-Cinca et al. 2015).

In contrast to the most relevant features for the default prediction, most of the variables measuring different income streams of the borrower were found to be among the least important

predictors in the classification models. Only the total income and the income from principal employer seem to have a notable effect on default probability in the Bondora dataset. Furthermore, even though the previous payment history was earlier found to be one of the most important determinants of default, the variables indicating the previous early repayments were found to be among the least important investigated features. Also, the variables indicating the employment status and employment duration of the borrower were ranked low in terms of feature importance estimates of the DT and RF classification models. Information about the most relevant and irrelevant features can be exploited by P2P platforms and researchers in the future when deciding which features should be considered when developing default prediction and credit scoring models in P2P lending context.

7 CONCLUSIONS

In this thesis, the performance of different supervised FS methods combined with different classification algorithms was investigated in P2P lending area. First, the basic principles of P2P lending were introduced, and the theoretical framework of the study was described. Then, the literature review of previous research was conducted. In the empirical part of the thesis, the performance of different FS methods was tested using the real-world dataset provided by an Estonian P2P lending platform Bondora. The performance of the tested methods was evaluated based on the final classification performance and model complexity.

According to the results of the empirical analysis, the SFS method outperformed other FS methods in P2P lending default prediction on Bondora dataset when measured by both final classification performance and reduction in model complexity. The SFS method was the only FS method that managed to improve the classification accuracy statistically significantly with almost all classification models compared to the model with the full feature set.

Even if improvements in the classification performance could not be achieved with all FS methods, the methods helped to reduce the number of features considerably without significantly reducing the classification performance. This helps to avoid excess complexity of the models and to improve their interpretability. It is also noteworthy that FS can considerably decrease the training time of classification algorithms which can be seen beneficial especially in the case of computationally heavy classification models.

The results of this study could be exploited by investors and P2P platforms when constructing the default prediction and credit scoring models for P2P lending. The study also offers insights on feature importance on new P2P lending dataset in default prediction area and demonstrates the performance of different FS methods on a new real-world application (the performance of

different FS methods has not earlier been compared systemically in P2P lending area). Thus, the results can also be exploited in the future research on the P2P lending field.

The limitations of this study must be considered when drawing conclusions from the results. The study was limited to a single dataset which limits the potential generalization of the results but was necessary to keep the scope of the thesis reasonable. Also, the scarceness of publicly available P2P lending datasets limited the use of multiple datasets in the analysis. It is also worth noting that the P2P lending markets are heterogenous by nature and the conclusions made based on one platform cannot necessarily be generalized to consider the whole industry.

The results are also dependent on the choice of used classification and FS methods. It is worth remarking that the used methods are just examples of the commonly applied classification and FS methods, and a lot of potential methods have been left out of this analysis. Investigating the performance of other FS and classification methods in P2P lending default prediction have been left to future research. In addition, it is worth noting that even though the classification performance was measured with multiple performance metrics, the statistical significance of results was tested only in terms of classification accuracy.

The ML models are computationally expensive and especially comprehensive parameter optimization and using wrapper-type FS methods are often intractable with a typical personal computer. Therefore, the use of more powerful computers and for example cloud computing should be considered in future research to enable a more comprehensive analysis. This would make it possible to ensure the optimality of used hyperparameters and therefore to make the results more reliable. The feature selection methods used in this study could also be tested with other P2P lending datasets.

REFERENCES

Abdou, H.A., Pointon, J. 2011. Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. Intelligent Systems in Accounting, Finance and Management, vol.18, no. 2, pp. 59-88.

Aksoy, S., Haralick, R.M. 2001. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern Recognition Letters, vol. 22, no. 5, pp. 563-582.

Almuallim, H, Dietterich, T. 1994. Learning Boolean concepts in the presence of many irrelevant features. Artificial Intelligence, vol. 69, no. 1, pp. 279-305.

Arlot, S., Celisse, A. 2010. A survey of cross-validation procedures for model selection. Statistics surveys, vol. 4, pp. 40-79.

Azim, T., Ahmed, S. 2018. Composing Fisher Kernels from Deep Neural Networks: A Practitioner's Approach. Springer Nature Switzerland AG.

Bachmann, A., Becker, A., Buerckner, D. 2011. Online Peer-to-Peer Lending – A Literature Review. Journal of Banking and Commerce, vol. 16, no. 2, pp. 1-18.

Bellotti, T., Crook, J. 2009. Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, vol. 36, no. 2, pp. 3302-3308.

Berger, S.C., Gleisner, F. 2009. Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending. BuR – Business Research, vol. 2, no. 1, pp. 39-65.

Bergstra, J., Bengio, Y. 2012. Random Search for Hyper-Parameter Optimization. Journal of machine learning research, vol. 13, pp. 281-305.

Bielecki, R., Rutkowski, M. 2004. Credit Risk: Modeling, Valuation and Hedging. Springer-Verlag Berlin Heidelberg, New York.

Bishop, C.M. 2006. Pattern recognition and machine learning. New York, Springer.

Blum, A.L., Langley, P. 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence, vol. 97, no. 1, pp. 245-271.

Bolon-Canedo, V., Sanchez-Marono, N., Alonso-Betanzos, A. 2013. Knowledge and Information Systems, vol. 34, pp. 483-519.

Bondora 2017. Background information about Bondora. Accessed 3.2.2020. Available https://support.bondora.com/hc/en-us/articles/212499589-Background-information-about-Bondora

Bondora 2019. Public Reports. Accessed 3.2.2020. Available https://www.bondora.com/en/public-reports

Bradley, A.P. 1997. The Use of the Area Uncer the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition, vol. 30, no. 7, pp. 1145-1159.

Breiman, L. 2001 Random Forests. Machine Learning, vol. 45, no. 1, pp. 5-32.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A.W., Siddique, A. 2016. Risk and risk management in the credit card industry. Journal of Banking and Finance, vol. 72, pp. 218-239.

Byanjankar, A., Heikkilä, M., Mezei, J. 2015. Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach. IEEE Symposium Series on Computational Intelligence, pp. 719-725.

Carmichael, D. 2014. Modeling default for peer-to-peer loans. Available at SSRN: http://ssrn.com/abstract=2529240, 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2529240.

Chang-Hwan, L., Gutierrez, F., Dejing, D. 2011. Calculating Feature Weights in Naïve Bayes with Kullback-Leibler Measure. 11th IEEE International Conference on Data Mining.

Chandrashekar, G., Sahin, F. 2014. A survey on feature selection methods. Computers and Electrical Engineering, vol. 40, no. 1, pp. 16-28.

Chen, W., Ma, C., Ma, L. 2009. Mining the customer credit using hybrid support vector machine technigue. Expert Systems with Applications, vol. 36, no. 4, pp. 7611-7616.

Chen, F.L., Li, F.C. 2010. Combination of feature selection approaches with SVM in credit scoring. Expert Systems with Applications, vol. 37, pp. 4902-4909.

Chen, D., Lai, F., Lin, X. 2014. A trust model for online peer-to-peer lending: a lender's perspective. Information Technology and Management, vol. 15, no. 4, pp. 239-254.

Chen, C.W.S., Dong, M.C., Liu, N., Scriboonchitta, S. 2019. Inferences of default risk and borrower characteristics on P2P lending. North American Journal of Economics and Finance, vol. 50.

Crook, J.N., Edelman, D.B., Thomas, L.C. 2007. Recent developments in consumer credit risk assessment. European Journal of Operational Research, vol. 183, no. 3, pp. 1447-1465.

Crotty, J. 2009. Structural causes of the global financial crisis: a critical assessment of the 'new financial architecture'. Cambridge Journal of Economics, vol. 33, no. 4, pp. 563-580.

Dahiya, S., Handa, S.S., Singh, N.P. 2017. A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. Expert Systems, vol. 34, no. 6.

Dash, M., Liu, H. 1997. Feature Selection for Classification. Intelligent Data Analysis, vol. 1, no. 1, pp. 131-156.

Dash, M., Liu, H. 2003. Consistency-based search in feature selection. Artificial Intelligence, vol. 151, pp. 155-176.

Davis, K., Murphy, J. (2016) Peer-to-Peer Lending: Structures, risks and regulation. JASSA: The Finsia Journal of Applied Finance, no. 3, pp. 37-44.

Dietterich, T.G. 1997a. Machine-Learning Research: Four Current Directions. AI Magazine, vol. 18, no. 4, pp. 97-136.

Dietterich T.G. 1997b. Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms. Neural Computation, vol. 10, no. 7, pp. 1895-1923.

Dietterich, T.G. 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization. Machine Learning, vol. 40, no. 2, pp. 139-157.

Ding, C., Peng H. 2003. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, vol. 3, no. 2, pp. 185–205.

Donders, A.R.T., Van Der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M. 2006. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiomology, vol. 59, no. 10, pp. 1087-1091.

Dorfleitner, G., Priberny, C., Schuster, S., Stoiber, J., Weber, M., de Castro, I., Kammler, J. 2016. Description-text related soft information in peer-to-peer lending – Evidence from two leading European platforms, vol. 64, pp. 169-187.

Dreiseitl, S., Ohno-Machado, L. 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, vol. 35, no. 5, pp. 352-359.

Duarte, R., Siegel, S., Young, L. 2012. Trust and Credit: The Role of Appearance in Peer-to-peer Lending. The Review of Financial Studies, vol. 25, no. 8, pp. 2455-2483.

Eunkyoung, L., Lee, B. 2012. Herding behavior in online P2P lending: An empirical investigation. Electronic Commerce Research and Applications, vol. 11, no. 5, pp. 495-503.

Emekter, R., Tu, Y., Jirasakuldech, B., Lu, M. 2015. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Applied Economics, vol. 47, no. 1, pp. 54-70.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874.

FinVolution Group 2019. FinVolution Group Reports Fourth Quarter and Fiscal Year 2019 Unaudited Financial Results and Announces Management Changes. Accessed 20.4.2020. Available https://ir.finvgroup.com/financial-reports

Freeman, C., Kulic, D., Basir, O. 2015. An evaluation of classifier-specific filter measure performance for feature selection. Pattern Recognition, vol. 48, no. 5, pp. 1812-1826.

Galindo, J., Tamayo, P. 2000. Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications, vol. 15, no. 1, pp. 107-143.

Guo, Y., Zhou, W., Luo, C., Liu, C., Xiong, H. 2016. Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, vol. 249, no. 2, pp. 417-426.

Guyon, I., Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, vol. 3, pp. 1157-1182.

Ha, V.S., Nguyen, H.N. 2016. An effective credit scoring model based on feature selection approaches. Proceedings of the First National Conference on Basic Research and Application of Information Technology (FAIR).

Hall, M. 1999. Correlation-based feature selection for machine learning. PhD thesis, Waikato University, Department of Computer Science.

Hart, C. (1998) Doing a Literature Review: Releasing the Social Science Research Imagination. SAGE Publications, London.

Huang, C., Chen, M., Wang, C. 2007. Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, vol. 33, no. 4, pp. 847-856.

Ishwaran, H. 2007. Variable importance in binary regression trees and forests. Electronic Journal of Statistics, vol. 1, pp. 519-537.

Iyer, R., Khwaja, A.I., Luttmer, E.F.P., Shue, K. 2009. Screening in New Credit Markets: Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending? AFA 2011 Denver Meeting Paper.

Japkowich, N., Shah, M. 2014. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, New York.

Jin, Y., Zhu, Y. 2015. A data-driven approach to predict default risk of loan for online Peer-to-Peer (P2P) lending. Fifth International Conference on Communication Systems and Network Technologies.

Kaufman, S., Rosset, S., Perlich, C., Stitelman, O. 2012. Leakage in Data Mining: Formulation, Detection and Avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 6, no. 4, pp. 1-21.

Kazemitabar, S.J., Amini, A.A., Bloniarz, A., Talwalkar; A. 2017. Variable Importance using Decision Trees. 31st Conference on Neural Information Processing Systems (NIPS).

Khalid, S., Khalil, T., Nasreen, S. 2014. A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. Science and information conference (SAI).

Khandani, A.E., Adlar, J.K., Lo, A.W. 2010. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, vol. 34, no. 11, pp. 2767-2787.

Khashman, A. 2010. Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications, vol. 37, no. 9, pp. 6233-6239.

Klafft, M. 2008. Online Peer-to-Peer Lending: A Lenders' Perspective. Proceedings of the International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, EEE 2008. H. R. Arabnia and A. Bahrami, eds., pp. 371-375, CSREA Press, Las Vegas 2008.

Kohavi, R., John, G.H. 1997. Wrappers for feature subset selection. Artificial Intelligence, vol. 97, no. 1-2, pp. 273-324.

Kotsiantis, S.B. 2007. Supervised Machine Learning: A Review of Classification Techniques. Informatica, vol. 31, no. 3, pp. 249-268.

Kruppa, J., Scwarz, A., Arminger, G., Ziegler, A. 2013. Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, vol. 40, no. 13, pp. 5125-5131.

Lal, T.N., Chapelle, O., Western, J., Elisseef, A. 2006. Embedded methods. Studies in Fuzziness and Soft Computing, vol. 207, pp. 137-165

Lee, T., Chiu, C., Chou, Y., Lu, C. 2006. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis, vol. 50, no. 4, pp. 1113-1130.

Lee, E., Lee, B. 2012. Herding behavior in online P2P lending: An empirical investigation. Electronic Commerce Research and Applications, vol. 11, no. 5, pp. 495-503.

Lending Club 2019. LendingClub Statistics. Accessed 10.2.2020. Available https://www.lendingclub.com/info/statistics.action

Lending Club 2020a. Interest Rates and Fees. Accessed 1.6.2020. Available https://www.lendingclub.com/investing/investor-education/interest-rates-and-fees

Lending Club 2020b. Rate information. Accessed 10.2.2020. Available https://www.lendingclub.com/foliofn/rateDetail.action

Liang, D., Tsai, C.F., Wu, H.T. 2015. The effect of feature selection on financial distress prediction. Knowledge-Based Systems, vol. 73, pp. 289-297.

Liberti, J.M., Petersen, M.A. 2018. Information: Hard and soft. IDEAS Working Paper Series from RePEc.

Lin, X., Li, X., Zheng, Z. 2017. Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China. Applied Economics, vol. 49, no. 35, pp. 3538-3545.

Liu, Y., Schumann, M. 2005. Data mining feature selection for credit scoring models. Journal of the Operational Research Society, vol. 56, no. 9, pp. 1099-1108.

Liu, H., Yu, S. 2005. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491-502.

Liu, H., Motoda, H. 2007. Computational Methods of Feature Selection. Boca Raton, CRC Press.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E. 2017. A survey of deep neural network architectures and their applications. Neurocomputing, vol. 234, pp. 11-26.

Loh, W.Y. 2002. “Regression Trees with Unbiased Variable Selection and Interaction Detection.” Statistica Sinica, vol. 12, pp. 361–386.

Louppe, G., Wehenkel, L., Sutera, A., Geurts, P. 2013. Understanding variable importances in forests of randomized trees. 27th Annual Conference on Neural Information Processing Systems (NIPS).

Louzada, F., Ara, A., Fernandes, G.B. Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science, vol. 21, no. 2, pp. 117-134.

Malekipirbazari, M., Aksakalli, V. 2015. Risk assessment in social lending via random forests. Expert Systems with Applications, vol. 42, no. 10, pp. 4621-4631.

Mantovani, R.G., Horvath, T., Cerri, R., Barbon Junior, S., Vanschoren, J., Carvalho, A.C.P.F. 2018. An empirical study on hyperparameter tuning of decision trees. arXiv:1812.02207.

McHugh, M.L. 2013. The Chi-Square test of independence. Biomechia Medica, vol. 23, no. 2, 143-149.

Michie, D. 1968. “Memo” functions and machine learning. Nature, vol. 218, no. 5136, pp. 19-22.

Mohri, M., Rostamizadeh, A., Talwalkar, A. 2012. Foundations of Machine Learning. Cambridge, MIT Press.

Onay, C., Ozturk, E. 2018. A review of credit scoring research in the age of Big Data. Journal of Financial Regulation and Compliance, vol. 26, no. 3, pp. 382-405.

Ong, C., Huang, J., Tzeng, G. 2005. Building credit scoring models using genetic programming. Expert systems with Applications, vol. 29, no. 1, pp. 41-47.

Oreski, S., Oreski, D., Oreski, G. 2012. Hybrid system with generic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Systems with Applications, vol. 39, no. 16, pp. 12605-12617.

Oreski, S., Oreski, G. 2014. Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Systems with Applications, vol. 41, no. 4, pp. 2052-2064.

P2PMarketData 2020. Top 80 Peer-2-Peer Lending & Equity by Funding Amounts. Accessed 20.4.2020. Available https://p2pmarketdata.com/

Patro, S.G., Sahu, K.K. 2015. Normalization: A Preprocessing Stage. arXiv:1503.06462.

Pelckmans, K., De Brabanter, J.D., Suykens, J.A.K., De Moor, B. 2005. Handling missing values in support vector machine classifier. Neural Networks, vol. 18, no. 5, pp. 684-692.

Peng, C.Y.J., Lee, K.L., Ingersoll, G.M. 2002. An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, vol. 96, no. 1, pp. 3-15.

Peng, H. Long, F., Ding, C. 2005. Feature selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238.

Pokorna, M., Sponer, M. 2016. Social lending and its risks. Procedia – Social and Behavioral Science, vol. 220, pp. 330-337.

Polena, M., Regner, T. 2018. Determinants of borrower's default in P2P lending under consideration of the loan risk class. Games, vol 9, no. 4.

Potdar, K., Pardawala, T.S., Pai, C.D. 2017. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications, vol. 175, no.4, pp. 7-9.

Powers, D.M.W. 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-63.

Probst, P., Wright, M.N., Boulesteix, A. 2019. Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery. Accessed 13.5.2020. Available https://doi.org/10.1002/widm.1301

Prosper 2020. Prosper Ratings. Accessed 20.3.2020. Available https://www.prosper.com/invest/how-to-invest/prosper-ratings/?mod=article_inline

Provost, F., Fawcett, T. 2013. Data Science for Business: What you need to know about data mining and data-analytic thinking. Sebastopol, O'Reilly Media.

Reunanen, J. 2003. Overfitting in making comparisons between variable selection methods. Journal of Machine Learning Research, vol. 3, pp. 1371-1382.

Saeys,Y., Inza, I., Larranaga, P. 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, vol. 23, no. 19, pp.2507-2517.

Safawian, S.R., Landgrebe, D. 1991. A survey on Decision Tree Classifier Methodology. IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674.

Sharmin, S., Ali, A.A., Khan, M.A.H., Shoyaib, M. 2017. Feature Selection and Discretization based on Mutual Information. IEEE International Conference on Imaging, Vision & Pattern Recognition 2017.

Somol, P. Baesens, B. Pudil, P., Vanthienen, J. 2005. Filter- versus Wrapper-based Feature Selection for Credit Scoring. International Journal of Intelligent Systems, vol. 20, no. 10, pp. 985-999.

Song, Y., Lu, Y. 2015. Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry, vol. 27, no. 2, pp. 130-135.

Serrano-Cinca, C., Gutierrez-Nieto, B., Lopez-Palacios, L. 2015. Determinants of default in P2P lending. PLoS One, vol. 10, no. 10.

Serrano-Cinca, C., Gutierrez-Nieto, B. 2016. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. Decision Support Systems, vol. 89, pp. 113-122.

Snoek, J., Larochelle, H., Adams, R.P. 2012. Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems, vol.4, pp. 2951-2959.

Suppers, A., Van Gool, A.J., Wessels, H. 2018. Integrated Chemometrics and Statistics to Drive Successful Proteomics Biomarker Discovery. Proteomes, vol. 6, no. 2.

Teply, P., Polena, M. 2020. Best classification algorithms in peer-to-peer lending. North American Journal of Economics and Finance, vol. 51.

Thomas, L.C. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, vol. 16, no. 2, pp. 149-172.

Tsai, C., Wu, J. 2008. Using neural network ensembles for bankruptcy prediction and credit scoring. Expert Systems and Applications, vol. 34, no. 4, pp. 2369-2649.

Van Gestel, T., Baesens, B. 2008. Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital. New York, Oxford University Press.

Wang, Y., Wang, S. and Lai, K.K. 2005. A New Fuzzy Support Vector Machine to Evaluate Credit Risk, IEEE Transactions on Fuzzy Systems, vol. 13, no. 6, pp. 820-831.

Wang, H., Chen, K., Zhu, W., Song, Z. 2015. A process model on P2P lending. Financial Innovation, vol. 1, no. 1, pp. 1-8.

Wang, D., Zhang, Z., Bai, R., Mao, Y. 2018. A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring. Journal of Computational and Applied Mathematics, vol. 329, pp. 307-321.

Webster, J., Watson, R.T. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, vol. 26, no. 2, pp. 13-23.

Wei, Z., Lin, M. (2017) Market Mechanisms in Online Peer-to-Peer Lending. Management Science, vol. 63, no.12, pp. 4236-4257.

West, D. 2000. Neural network credit scoring models. Computers & Operations Research, vol. 27, no. 8, pp. 1131-1152.

Xia, Y., Liu, C., Liu, N. 2017. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electronic Commerce Research and Applications, vol. 24, pp. 30-49.

Ye, X., Dong, L., Ma, D. 2018. Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score, Electronic Commerce Research and Applications, vol. 32, pp. 23-36.

Yeh, I., Lien, C. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, vol. 36, no. 2, pp. 2473-2480.

Yle 2019. Pikavippimarkkinat muuttuvat nyt, ja tässä ovat seuraukset: lainansaanti vaikeutuu, maksuhäiriöt lisääntyvät – monen velkakierre voi myös katketa. Accessed 19.3.2020. Available https://yle.fi/uutiset/3-10943120.

Yu, L., Liu, H. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research, vol. 5, pp. 1205-1224.

Yu, L., Wang, S., Lai, K.K. 2008. Credit risk assessment with a multistage neural network ensemble learning approach. Expert Systems with Applications, vol. 34, no. 2, pp. 1434-1444.

Yuang, J., Wang, J.G. 2016. Financing Without Bank Loans: New Alternatives for Funding SMEs in China. Singapore, Springer.

Yum, H., Lee, B., Chae, M. 2012. From the wisdom of crowds to my own judgement in microfinance through online peer-to-peer lending platforms. Electronic Commerce and Applications, vol. 11, no. 5, pp. 469-483.

Zhang, H. 2005. Exploring conditions for the optimality of Naïve Bayes. International Journal of Pattern Recognition and Artificial Intelligence, vol. 19, no. 2, pp. 183-198.

Zhang, Y., Jia, H., Diao, Y., Hai, M., Li, H. 2016. Research on Credit Scoring by fusing social media information in Online Peer-to-Peer Lending. Procedia Computer Science, vol. 91, pp. 168-174.

Zhao, H., Ge, Y., Liu, Q., Wang, G., Chen, E., Zhang, H. 2017. P2P Lending Survey: Platforms, Recent Advances and Prospects. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 8, no. 6, pp. 1-28.

Zheng, Z., Wu, X., Srihari, R. 2004. Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 80-89.

Zhou, L., Lai, K.K., Yu, L. 2010. Least squares support vector machines ensemble models for credit scoring. Expert Systems with Applications, vol. 37, no. 1, pp. 127-133.

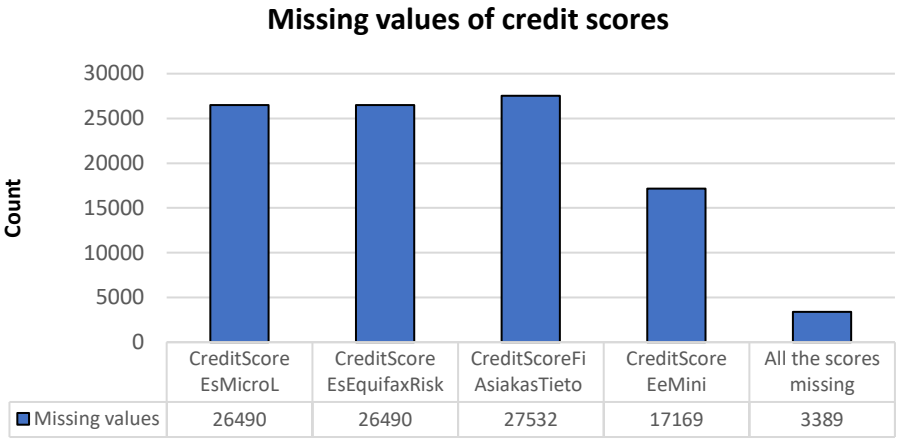
Zhou, J., Li, W., Wang, J., Ding, S., Xia, C. 2019. Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A: Statistical Mechanics and its Applications, vol. 534

APPENDICES

Appendix 1. Main objectives and used data of reviewed studies from credit risk area

Author(s) & year	Main objective	Data
Wang et al. 2018	Comparing the classification performance of ensemble learning methods in credit scoring	Australian credit dataset, German credit dataset and a dataset from Commercial Bank of China
Dahiya et al. 2017	Introducing a new hybrid-bagging classification algorithm with FS for credit risk evaluation	German credit dataset
Butaru et al. 2016	Examining the credit risk management practices in financial institutions	Credit card data from several large US financial institutions
Ha and Nguyen 2016	Using FS to enhance the performance of credit scoring	Australian credit dataset and German credit dataset
Louzada et al. 2016	Reviewing the classification methods in credit scoring	Previous studies related to the topic
Oreski and Oreski 2014	Presenting a novel FS algorithm for credit risk assessment	Credit dataset from a Croatian bank, German credit dataset
Kruppa et al. 2013	Using ML methods to assess the probability of default in consumer credit area	Dataset from household appliances company offering payment by installment
Oreski et al. 2012	Enhancing the retail credit risk assessment with FS and hybrid classification system	Credit dataset from a Croatian bank
Abdou 2011	Reviewing both traditional and sophisticated statistical techniques and evaluation criteria in credit scoring	214 articles, books and theses involving credit scoring applications
Chen and Li 2010	Combining different FS approaches with SVM for credit scoring	Australian credit dataset, German credit dataset
Khandani et al. 2010	Using ML techniques for forecasting the consumer credit risk	Bank's dataset consisting of both transaction-level, credit bureau and account-balance data for individual consumers
Khashman 2010	Comparing the performance of NN models with different learning schemes in credit risk evaluation	German credit dataset
Zhou et al. 2010	Using SVM-based ensemble methods for credit scoring	German credit dataset, dataset from an English financial services company
Bellotti and Crook 2009	Using SVM for credit scoring and exploring the significant features	Credit card dataset
Chen et al. 2009	Introducing hybrid SVM model with FS for credit scoring	Credit card dataset from Chinese local bank
Yeh and Lien 2009	Comparing the predictive accuracy of different data mining models in credit card client default probability prediction	Payment data from a big Taiwanese bank
Yu et al. 2008	Introducing a NN ensemble model for credit risk assessment	Japanese consumer credit card application data
Tsai and Wu 2008	Using NN ensemble for bankruptcy prediction and credit scoring	Australian credit dataset, German credit dataset and Japanese credit dataset
Huang et al. 2007	Using hybrid SVM model for credit scoring	Australian credit dataset and German credit dataset
Lee et al. 2006	Using DT and adaptive regression splines for credit scoring	Credit card dataset from a Taiwanese local bank
Liu and Schumann 2005	Using FS methods to enhance the performance of credit scoring models	Dataset from a German credit insurance company
Ong et al. 2005	Using genetic programming for credit scoring	Australian credit dataset, German credit dataset
Somol et al. 2005	Comparing the filter-based and wrapper-based FS methods in credit scoring	Australian credit dataset, German credit dataset and two datasets from major financial institutions in Benelux countries
Wang et al. 2005	Comparing the performance of different ensemble learning methods in credit scoring	Australian credit dataset, German credit dataset and Chinese credit dataset
Galindo and Tamayo 2000	Using statistical and ML models for credit risk assessment	Mortgage loan data from Mexico's security exchange and banking commission
Thomas 2000	Reviewing the credit and behavioral scoring techniques	Previous literature related to the topic
West 2000	Investigating the credit scoring accuracy of five different NN models	Australian credit dataset, German credit dataset

Appendix 2. Missing values of different credit scores



Appendix 3. The descriptions of used features

Group and variable	Definition
Credit risk assessment	
Credit rating	The credit rating assigned by Bondora
Credit score	The credit score assigned by third party
Interest rate	Maximum interest accepted by borrower in the loan application
Demographics of the borrower	
Language	Language of the borrower
Gender	Gender of the borrower
Country	Borrower residency
Age	The age of the borrower in years when signing the loan application
Month of birth	The month the borrower has been born in
Home ownership type	Type of the ownership of the borrower's home
Education	Educational level of the borrower
Employment status	Employment status of the borrower
Employment duration current employer	The length of employment with current employer
Work experience	Overall work experience of the borrower
Occupation area	Occupational area of the borrower
Marital status	Marital status of the borrower
Number of dependants	Number of children and other dependants of the borrower
Income information	
Income from principal employer	The income of the borrower from the current employer
Income from pension	The income of the borrower from pension
Income from family allowance	The income of the borrower from child support
Income from social welfare	The income of the borrower from social support
Income from leave pay	The income of the borrower from paternity leave
Income from child support	The income of the borrower from alimony payments
Income other	Other income of the borrower
Total income	The total monthly income of the borrower
Characteristics of the loan	
Applied amount	The loan amount the borrower originally applied for
Loan duration	The loan duration in months
Use of loan	The use of the loan reported by borrower
Monthly payment day	The day of month when the monthly payments are made
Application signed hour	The hour when the application was signed
Application signed weekday	The weekday the application was signed
Loan month	The month the loan was issued

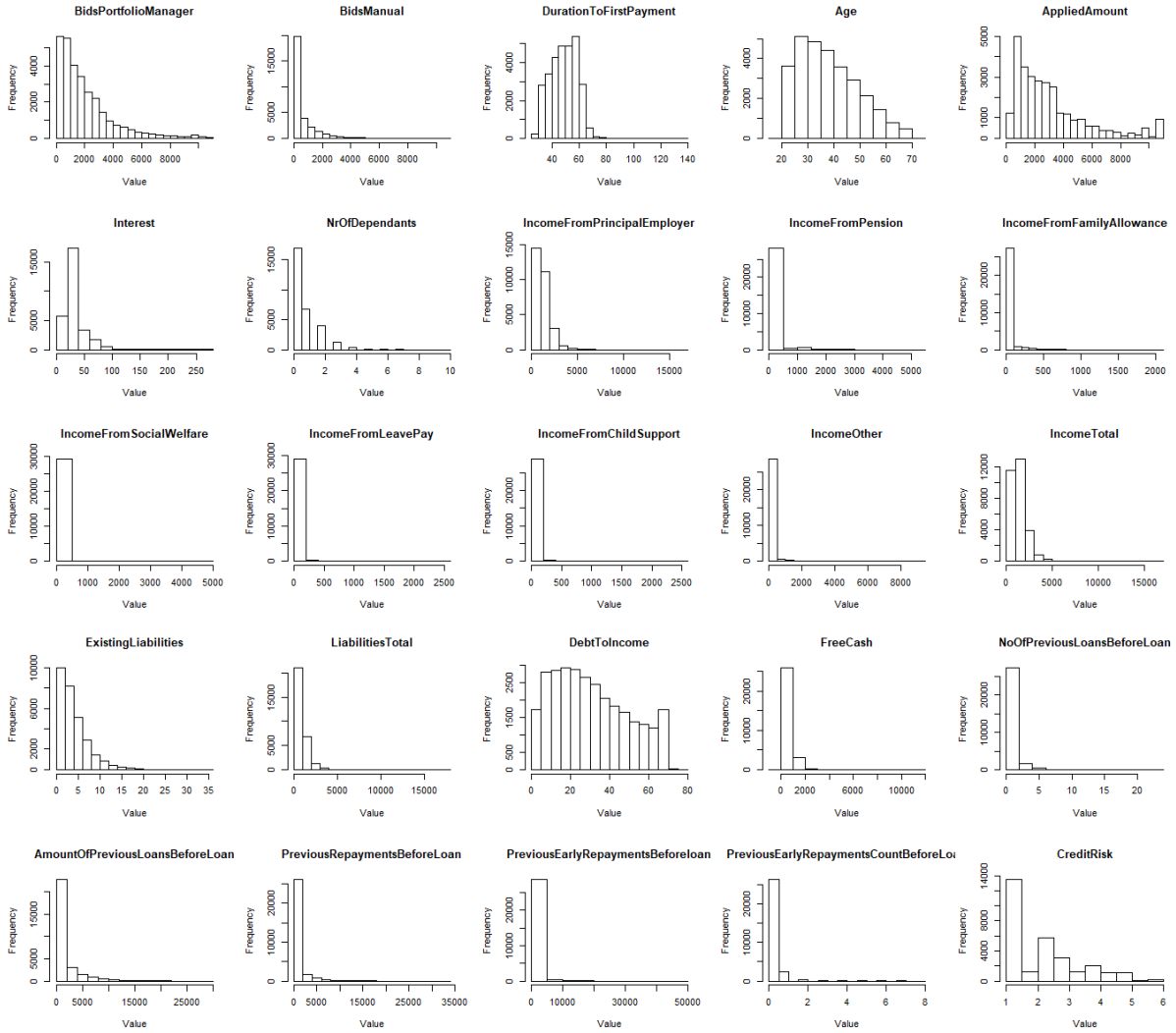
Duration to first payment	Duration to the first payment of the loan, calculated as the difference between loan application day and the first scheduled payment day
Verification type	The method used to verify the loan application data
Credit history of the borrower	
New credit customer	Whether the customer has prior credits in Bondora or not
Number of previous loans	Number of previous loans before the loan being applied
Amount of previous loans	Amount of previous loans before the loan being applied
Amount of previous repayments	Amount of previous repayments before the loan being applied
Number of previous early repayments	Number of early repayments before the loan being applied
Amount of previous early repayments	Amount of early repayments before the loan being applied
Indebtedness of the borrower	
Existing liabilities	Number of existing liabilities of the borrower
Liabilities total	The amount of total liabilities of borrower
Debt to income ratio	The ratio of debt to gross monthly income of the borrower
Free cash	Free cash after subtraction of monthly liabilities
Others	
Bids portfolio manager	The amount of loan offers made by Portfolio Manager
Bids manual	The amount of loan offers borrower had made manually

Source: Bondora 2019. Public Reports. Accessed 3.2.2020. Available https://www.bondora.com/en/public-reports

Appendix 4. Summary statistics of categorical features

Variable	Min.	Max.	Mean	Median	STD	Skewness	Kurtosis
Bids portfolio manager	0.00	10625.00	1984.33	1460.00	1926.38	1.79	6.72
Bids manual	0.00	10630.00	563.58	25.00	990.64	3.23	18.87
Duration to first payment	28.00	136.00	49.06	49.00	9.57	0.00	2.83
Age	19.00	72.00	38.38	37.00	11.38	0.56	2.55
Applied amount	100.00	10630.00	3085.03	2230.00	2561.56	1.41	4.43
Interest rate	6.00	263.63	35.84	30.13	27.25	3.96	24.53
Number of dependants	1.00	10.00	3.02	2.00	1.72	1.49	4.01
Income from principal employer	0.00	17000.00	1198.85	1028.00	828.34	2.48	26.92
Income from pension	0.00	5038.00	74.82	0.00	296.77	5.38	36.76
Income from family allowance	0.00	2006.00	23.46	0.00	75.50	5.99	63.39
Income from social welfare	0.00	4551.00	10.11	0.00	76.26	19.52	714.23
Income from leave pay	0.00	2500.00	11.44	0.00	97.92	11.62	169.66
Income from child support	0.00	2500.00	9.83	0.00	57.85	9.89	193.56
Income other	0.00	9200.00	56.19	0.00	257.87	11.15	226.16
Income total	200.00	17000.00	1384.72	1200.00	824.55	2.92	29.26
Existing liabilities	0.00	36.00	4.46	4.00	3.34	1.70	7.33
Liabilities total	0.00	17435.00	855.81	675.00	644.43	3.42	35.34
Debt to income	0.00	79.96	30.93	28.09	19.05	0.41	2.12
Free cash	-76.93	11508.11	469.79	360.80	477.62	3.31	40.60
Number of previous loans	0.00	23.00	0.65	0.00	1.22	4.02	38.02
Amount of previous loans	0.00	30000.00	1349.72	0.00	2656.38	2.78	12.75
Number of previous repayments	0.00	33874.18	698.75	0.00	1823.19	4.74	35.54
Amount of previous early repayments	0.00	48100.00	332.44	0.00	1544.48	9.28	154.51
Number of previous early repayments	0.00	8.00	0.13	0.00	0.48	5.51	49.33
Credit risk	1.00	6.00	2.11	2.00	1.22	0.83	2.81

Appendix 5. Distributions of numerical features



Appendix 6. Class frequencies of categorical predictors

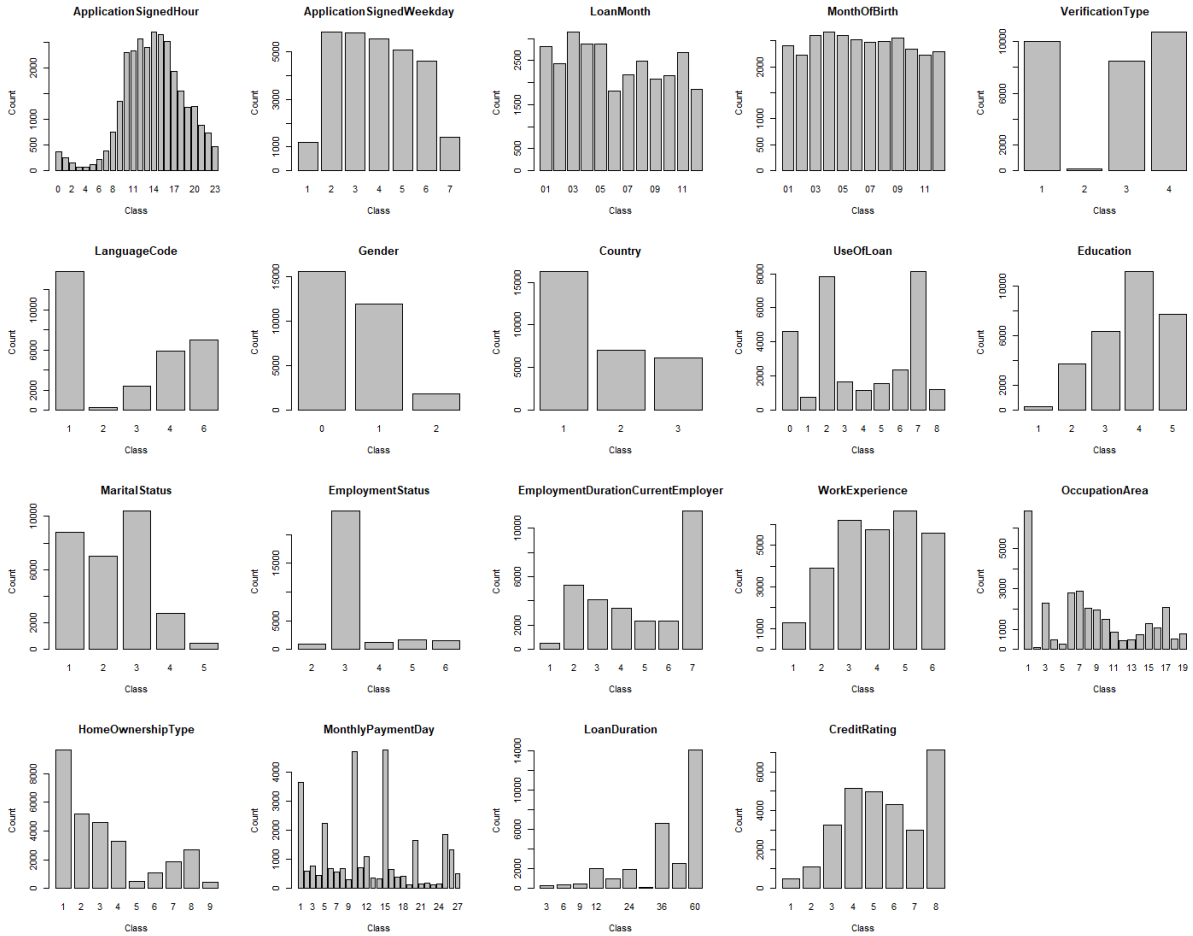
Variable	Class	Frequency	Relative frequency
New credit customer	**0 = False**	10240	34.86%
	1 = True	19135	65.14%
Verification type	**1 = Income unverified**	10026	34.13%
	2 = Income unverified, cross-referenced by phone	178	0.61%
	3 = Income verified	8474	28.85%
	4 = Income and expenses verified	10697	36.42%
Language	**1 = Estonian**	13777	46.90%
	2 = English	263	0.90%
	3 = Russian	2424	8.25%
	4 = Finnish	5924	20.17%
	6 = Spanish	6987	23.79%
Gender	**0 = Male**	15537	52.89%
	1 = Female	11956	40.70%
	2 = Unknown	1882	6.41%
Country	**1 = Estonia**	16219	55.21%
	2 = Spain	7034	23.95%
	3 = Finland	6122	20.84%
Use of loan	**0 = Loan consolidation**	4628	15.75%

	1 = Real estate	789	2.69%
	2 = Home improvement	7831	26.66%
	3 = Business	1689	5.75%
	4 = Education	1181	4.02%
	5 = Travel	1558	5.30%
	6 = Vehicle	2384	8.12%
	7 = Other	8105	27.59%
	8 = Health	1210	4.12%
Education	**1 = Primary education**	323	1.10%
	2 = Basic education	3780	12.87%
	3 = Vocational education	6357	21.64%
	4 = Secondary education	11177	38.05%
	5 = Higher education	7738	26.34%
Marital status	**1 = Married**	8781	29.89%
	2 = Cohabitant	7041	23.97%
	3 = Single	10380	35.34%
	4 = Divorced	2735	9.31%
	5 = Widow	438	1.49%
Employment status	**2 = Partially unemployed**	952	3.24%
	3 = Fully employed	24078	81.97%
	4 = Self-employed	1170	3.98%
	5 = Entrepreneur	1712	5.83%
	6 = Retiree	1463	4.98%
Employment duration current employer	**1 = Trial period**	498	1.70%
	2 = Up to 1 year	5316	18.10%
	3 = Up to 2 years	4111	13.99%
	4 = Up to 3 years	3371	11.48%
	5 = Up to 4 years	2341	7.97%
	6 = Up to 5 years	2340	7.97%
	7 = More than 5 years	11398	38.80%
Work experience	**1 = Less than 2 years**	1282	4.36%
	2 = 2 to 5 years	3913	13.32%
	3 = 5 to 10 years	6198	21.10%
	4 = 10 to 15 years	5757	19.60%
	5 = 15 to 25 years	6637	22.59%
	6 = More than 25 years	5588	19.02%
Occupation area	**1 = Other**	6835	23.27%
	2 = Real estate	111	0.38%
	3 = Research	2308	7.86%
	4 = Administrative	492	1.67%
	5 = Civil service and military	270	0.92%
	6 = Education	2807	9.56%
	7 = Healthcare and social help	2879	9.80%
	8 = Art and entertainment	2023	6.89%
	9 = Agriculture, forestry and fishing	1972	6.71%
	10 = Mining	1487	5.06%
	11 = Processing	861	2.93%
	12 = Energy	439	1.49%
	13 = Utilities	490	1.67%
	14 = Construction	724	2.46%
	15 = Retail and wholesale	1293	4.40%
	16 = Transport and warehousing	1050	3.57%
	17 = Hospitality and caring	2064	7.03%
	18 = Info and telecom	510	1.74%
	19 = Finance and insurance	760	2.59%
Home ownership type	**1 = Owner**	9629	32.78%
	2 = Living with parents	5221	17.77%
	3 = Tenant, pre-furnished property	4599	15.66%

	4 = Tenant, unfurnished property	3287	11.19%
	5 = Council house	501	1.71%
	6 = Joint tenant	1102	3.75%
	7 = Joint ownership	1875	6.38%
	8 = Mortgage	2722	9.27%
	9 = Owner with encumbrance	439	1.49%
Loan duration	**3 = 3 months**	314	1.07%
	6 = 6 months	333	1.13%
	9 = 9 months	443	1.51%
	12 = 12 months	2040	6.94%
	18 = 18 months	1014	3.45%
	24 = 24 months	1917	6.53%
	30 = 30 months	70	0.24%
	36 = 36 months	6651	22.64%
	48 = 48 months	2545	8.66%
	60 = 60 months	14048	47.82%
Credit rating	**1 = AA (the safest grade)**	490	1.67 %
	2 = A	1098	3.74 %
	3 = B	3270	11.13 %
	4 = C	5140	17.50 %
	5 = D	4974	16.93 %
	6 = E	4316	14.69 %
	7 = F	2980	10.14 %
	8 = HR (the riskiest grade)	7107	24.19 %
All the variables	**Overall**	29375	100.0 %

Note: To keep the table simple, variables that indicate dates are excluded.

Appendix 7. Distributions of categorical predictors



Appendix 8. Point-biserial correlations (continuous predictors and target)

Variable	Correlation coefficient	t-value	p-value
Bids portfolio manager	0.0281	4.889	1.02E-06
Bids manual	0.0590	10.275	9.99E-25
Duration to first payment	0.0401	6.983	2.95E-12
Age	-0.0062	-1.083	0.279
Applied amount	0.0954	16.673	3.93E-62
Interest rate	**0.2287**	**40.883**	**0.00E+00**
Number of dependants	-0.0220	-3.831	0.000128
Income from principal employer	0.0908	15.862	1.97E-56
Income from pension	0.0447	7.790	6.91E-15
Income from family allowance	-0.0007	-0.126	0.0900
Income from social welfare	0.0219	3.819	0.000134
Income from leave pay	-0.0180	-3.125	0.00178
Income from child support	-0.0044	-0.761	0.4470
Income other	-0.0258	-4.493	7.05E-06
Income total	0.0988	17.267	1.75E-66
Existing liabilities	0.0032	0.564	0.573
Liabilities total	0.0784	13.678	1.85E-42
Debt to income	0.0282	4.901	9.59E-07
Free cash	0.0659	11.493	1.66E-30
Number of previous loans	**-0.1431**	**-25.153**	**3.48E-138**
Amount of previous loans	**-0.1129**	**-19.763**	**2.20E-86**
Number of previous repayments	**-0.1522**	**-26.792**	**2.66E-156**
Amount of previous early repayments	-0.0122	-2.121	0.0339
Number of previous early repayments	-0.0264	-4.603	4.17E-06
Credit score	**0.2548**	**45.837**	**0.00E+00**

The largest absolute correlation coefficient values are bolded.

Appendix 9. Chi-Square test of independence (categorical predictors and target)

Variable	Chi-Square test statistic	p-value
New credit customer	**644.03**	**4.44E-142**
Application signed hour	102.19	5.88E-12
Application signed weekday	14.99	0.020315
Loan month	81.44	7.77E-13
Month of birth	26.13	0.0062
Verification type	305.92	5.21E-66
Language code	246.03	4.67E-52
Gender	**638.79**	**1.94E-139**
Country	**3114.02**	**0**
Use of loan	116.96	1.40E-21
Education	**632.28**	**1.59E-135**
Marital status	179.15	1.13E-37
Employment status	246.03	4.67E-52
Employment duration current employer	10.78	0.095459
Work experience	25.05	0.000136
Occupation area	324.84	3.65E-58
Home ownership type	472.36	5.97E-97
Monthly payment day	**1034.96**	**1.44E-201**
Loan duration	**697.68**	**2.18E-144**
Credit rating	**3461.08**	**0**

The largest Chi-Square test statistic values are bolded.

Appendix 10. Class frequencies of target variable across categorical variables

	Whole data		Training data		Test data		
Feature	**Non-default**	**Default**	**Non-default**	**Default**	**Non-default**	**Default**	**% of whole data**
Credit rating							
1 = AA	89.6%	10.4%	89.1%	10.9%	90.8%	9.2%	1.67%
2 = A	76.8%	23.2%	76.2%	23.8%	78.0%	22.0%	3.74%
3 = B	66.8%	33.2%	67.4%	32.6%	65.4%	34.6%	11.13%
4 = C	58.0%	42.0%	57.8%	42.2%	58.6%	41.4%	17.50%
5 = D	44.1%	55.9%	44.5%	55.5%	43.0%	57.0%	16.93%
6 = E	36.6%	63.4%	36.8%	63.2%	36.1%	63.9%	14.69%
7 = F	34.8%	65.2%	34.6%	65.4%	35.3%	64.7%	10.14%
8 = HR	23.0%	77.0%	22.8%	77.2%	23.5%	76.5%	24.19%
Country							
1 = Estonia	58.3%	41.7%	58.3%	41.7%	58.3%	41.7%	55.21%
2 = Spain	23.1%	76.9%	22.9%	77.1%	23.7%	76.3%	23.95%
3 = Finland	29.5%	70.5%	29.8%	70.2%	29.0%	71.0%	20.84%

Appendix 11. In-sample and 5-fold CV errors for different NB models

Model	**In-sample error**	**5-fold CV error**
NB (No FS)	0.333	0.335
NB + MRMR	0.321	0.322
NB + Chi-Square	0.327	0.328
NB + SFS	0.319	0.320

Appendix 12. In-sample and 5-fold CV errors for different LR models

Model	**In-sample error**	**5-fold CV error**
LR (No FS)	0.291	0.299
LR + MRMR	0.296	0.302
LR + Chi-Square	0.296	0.305
LR + SFS	0.302	0.298

Appendix 13. In-sample and 5-fold CV errors for different DT models

	Default hyperparameters		**Optimized hyperparameters**	
Model	**In-sample error**	**5-fold CV error**	**In-sample error**	**5-fold CV error**
DT (No FS)	0.049	0.393	0.314	0.317
DT + MRMR	0.098	0.390	0.304	0.316
DT + Chi-Square	0.248	0.380	0.309	0.317
DT + SFS	0.113	0.350	0.307	0.312
DT + LMBFR	0.101	0.385	0.316	0.318

Appendix 14. 5-fold CV errors for different RF models

	Default hyperparameters	**Optimized hyperparameters**
Model	**5-fold CV error**	**5-fold CV error**
RF (No FS)	0.293	0.297
RF + MRMR	0.298	0.299
RF + Chi-Square	0.291	0.291
RF + SFS	0.271	0.295
RF + LMBFR	0.286	0.300